INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Xerox University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48108
ANDERSON, Deborah Jean, 1943-
AN EVALUATION OF THE METHODOLOGICAL AND POLICY IMPLICATIONS OF THE D.C. POLICEWOMEN ON PATROL STUDY.

The Ohio State University, Ph.D., 1976
Psychology, clinical

Xerox University Microfilms, Ann Arbor, Michigan 48106

Copyright by
Deborah Jean Anderson
1976
AN EVALUATION OF THE METHODOLOGICAL AND POLICY IMPLICATIONS
OF THE D.C. POLICEWOMEN ON PATROL STUDY

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

By
Deborah Jean Anderson, B.A., M.A.

* * * * *

The Ohio State University
1976

Reading Committee:
Jaques Kaswan, Ph.D.
Saul Siegel, Ph.D.
Andrew Schwebel, Ph.D.

Approved By
Adviser
Department of Psychology
DEDICATION

To my father, who valued the search for understanding, who saw no barriers as impossible, and who wanted his daughters to be free.
ACKNOWLEDGEMENTS

First, I am very grateful to Dr. Jaques Kaswan, my advisor in this venture, for a great many things—his understanding of the personal implications and choices at all times, his encouragement and support throughout, his wisdom, guidance and foresight in following (or preced-eding) me on difficult and at times dimly lit paths, his sensitivity to personal and professional issues, and his consistent generosity of time and spirit whenever needed. The experience of conducting this disser-tation, although lengthy, was far more instructive and stimulating than could normally have been expected, which in large part is due to the flexibility, excellent advising and friendship given by Dr. Kaswan.

I must also thank Peter Bloch, who gave me the chance to leave the student world, and to participate in a study which interested us both very much. The experience was invaluable in learning about research, and I am grateful for his friendship and for the opportunity to work closely with someone who is so committed to "making a difference." Saul Siegel I can thank for many good years at this institution, for encouraging new ways, and for a warmth which is very much appreciated. In addition, his encouragement of my sojourn in Washington, as a step towards this dissertation, as well as his being a member of the reading committee, has been personally and professionally helpful. To Andy Schwebel, as a member of my reading committee, who was helpful in the
planning stages, with his interest in and knowledge of this area, I again give thanks.

A number of other people have helped in innumerable ways, without whom this effort would have been considerably more difficult. My family, in particular, my mother, was extremely supportive and generous throughout. My sister and brother helped ease the bleak days in Minnesota, and gave a warmth not often found. Caroline and Ami could be counted on to help in the worst hours and days, as well as the best. Montina helped with the research and shared the joy of living. Ruth Kaswan lent her support and home as well. Dr. Jack Freimeier, the graduate representative, was kind enough to be interested in the study. Patti Watson rescued me time and again with the monumental typing involved, well beyond what ever should have been expected. Ruth and Karen came in at the eleventh hour to make sure this would be finished. And others—too many friends to be mentioned—helped me find a warm place again in an old place.
VITA

March 11, 1943 .................. Born—Leavenworth, Kansas

1965 ............................ B.A., University of Minnesota, Minneapolis, Minnesota

1967 ............................ M.A., The Ohio State University, Columbus, Ohio

1967-1968 ....................... Clinical internship, Mt. Zion Hospital, San Francisco

1968-1969 ....................... Teaching assistant, research assistant/associate, The Ohio State University, Columbus, Ohio

1969-1971 ....................... Staff psychologist, Columbus State Hospital, Columbus, Ohio

1972-1975 ....................... Research Associate I, The Urban Institute, Washington, D.C.

PUBLICATIONS


FIELDS OF STUDY

Major Field: Clinical Psychology

Studies in Community Psychology. Professors Jaques Kaswan and James Kelly.

Studies in Psychopathology. Professors Saul Siegel and Thomas Weaver.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEDICATION</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>VITA</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>ix</td>
</tr>
<tr>
<td>Chapter</td>
<td></td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>II. METHODOLOGY</td>
<td>33</td>
</tr>
<tr>
<td>III. RESULTS</td>
<td>102</td>
</tr>
<tr>
<td>IV. DISCUSSION</td>
<td>136</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>169</td>
</tr>
<tr>
<td>A.</td>
<td>170</td>
</tr>
<tr>
<td>B.</td>
<td>182</td>
</tr>
<tr>
<td>C.</td>
<td>215</td>
</tr>
<tr>
<td>D.</td>
<td>253</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>265</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Characteristics of experimental and comparison police districts during fiscal year 1972.</td>
<td>38</td>
</tr>
<tr>
<td>2. Assignments of first year study samples after one year.</td>
<td>44</td>
</tr>
<tr>
<td>3. Measurement instruments.</td>
<td>46</td>
</tr>
<tr>
<td>4. Dispatched incidents per tour.</td>
<td>108</td>
</tr>
<tr>
<td>5. Misdemeanor arrests during first six months of 1973.</td>
<td>115</td>
</tr>
<tr>
<td>6. Prosecutor's data on arrests and dispositions for comparison men and new women.</td>
<td>117</td>
</tr>
<tr>
<td>7. Police data on arrests and dispositions for comparison men and new women.</td>
<td>118</td>
</tr>
<tr>
<td>8. Departmental performance ratings.</td>
<td>124</td>
</tr>
<tr>
<td>9. Percent of officer with one year's experience believed &quot;satisfactory&quot; at patrol skills.</td>
<td>129</td>
</tr>
<tr>
<td>10. Chief's Survey Sample--1972.</td>
<td>220</td>
</tr>
<tr>
<td>11. Official's Survey Ratings Sample</td>
<td>223</td>
</tr>
<tr>
<td>12. Level of patrol activity.</td>
<td>227</td>
</tr>
<tr>
<td>13. Service Survey Sample characteristics.</td>
<td>234</td>
</tr>
<tr>
<td>14. Community and business survey samples.</td>
<td>239</td>
</tr>
<tr>
<td>15. Patrol Survey Sample.</td>
<td>244</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>216</td>
</tr>
<tr>
<td>Figure 2: Interviewer's Grid</td>
<td>241</td>
</tr>
</tbody>
</table>
INTRODUCTION

In 1972, the Metropolitan Police Department of the District of Columbia initiated an experimental program to determine whether women could be used interchangeably with men on uniformed patrol duty. This project was evaluated by the Urban Institute, during which time the author was involved in all aspects of the design and conduct of the evaluation, as study co-author. The study was designed to assess the ability of women, in comparison with men of similar experience, to perform routine patrol functions. The results from this study have implications for the larger issue of sex discrimination within police departments.

STATEMENT OF THE PROBLEM

The purpose of this dissertation is to examine the evaluation of the policewomen program in Washington, D.C. Since the author was directly involved with this research for the two year duration, this provides an opportunity to view the ongoing process and problems involved in conducting such research with knowledge of the real constraints and reasons for the procedures taken. A number of issues which are central to the conduct of field research, yet which are not generally known to those involved in critiquing such research, can be explored more thoroughly and objectively as a dissertation effort. They include
the following: the processes involved in the generation and conduct of such research; the reasons for selection of the particular measures and methodological strategies which were used; the strengths and weaknesses of such approaches; the issues involved in the interpretation and function of results; and the factors which influenced the utilization of this evaluation. The influence of the different goals and values of the organizations and persons involved in the evaluation, the policy implications of this research, and the role of evaluators and of evaluation research in social change efforts will also be discussed. Examination of these issues, as they pertain to the policewomen study and in their broader application to field research efforts, can contribute to the development of more sophisticated field research strategies.

This study was selected as the focus for a dissertation for a variety of reasons. It provides an example of field research in which considerable effort was expended to insure that the design, measurement instruments, and analytical techniques were as rigorous as possible, given the tight time, staff and financial constraints. In this sense, this study illustrates many of the methodological and policy advantages and disadvantages of field research. The current literature dealing with field research tends to be polarized between the methodological or political critiques of the academicians, who generally have little practical experience with the realities of such research, and the more practically oriented writings of those directly involved in such evaluation efforts, who tend to shun exploration of the assumptions upon which they are operating, and, for reasons of
expedience or lack of expertise, often perform evaluations which are methodologically quite deficient.

DEFINITIONS OF EVALUATION RESEARCH

The evaluation of public programs is a highly topical and controversial issue. The design of this study may be clarified by a brief review of the different types of evaluation research which have been used or proposed.

Field vs. Laboratory Research

Field research, as distinguished from laboratory research, generally is conducted in order to provide data for administrative or policy decisions, whereas the purpose of most laboratory research is to contribute to the field of knowledge, with often scant attention being paid to the practical application of the results. Laboratory research is often performed within the framework of particular theoretical models, each area of inquiry often having methodological strategies and measurement instruments which are commonly employed. Generally, either the theoretical constructs are tested, using instruments which have been validated and standard methodological approaches, or the instruments and/or strategies are examined with respect to reliability and validity. A data base of previous studies in the area is often available to suggest relevant questions and approaches to the problem.

Since field research is conducted in natural rather than controlled settings, the methodology is generally less rigorous than with laboratory research. The reliability and validity of the measurement instruments is generally unknown, since there is often little or no
data from similar studies to draw upon, and there is rarely a theoretical model to guide the questions asked. In addition, there are usually many factors external to the program which influence the evaluation results, but which are often difficult or impossible to control for or even assess. Hence, field research is less likely than laboratory research to adequately control for alternative explanations of effects. Finally, elucidating in advance the exact procedures which will be applied in any program is not always possible, which complicates the process of designing adequate evaluations. Thus, there is a larger area of uncertainty and unreliability in field research than in the academic variety.

The populations utilized in laboratory research are often selected on the basis of availability rather than relevance. In field research, program participants are often drawn from the populations deemed most in need of such services, rather than randomly selected, which at times creates difficulties in selecting adequate control populations. In some cases, programs are applied to the entire eligible population, leaving no subjects as controls. In any case, the study population is generally a subsample of the actual population about which decisions are to be made. The conditions under which field and laboratory research are conducted, as well as the populations which are studied, have given rise to the polarized conception of field research as "natural" and laboratory research as "artificial" (see Festinger and Katz, 1953 for a thorough discussion of these issues).
Because the major goal of field research is to provide information for use in decision-making, the costs and benefits of making such decisions must be weighed against the costs and benefits of failing to make such decisions, given data with less than optimum validity and reliability. If there is strong commitment to the amelioration of certain problems, then less stringent evidence of a desired effect is necessary than in a purely academic study. For example, in the policewomen evaluation, evidence of a sizeable overlap in the performance of male and female police officers was sufficient to support the program of women on patrol, even if group differences occurred which favored the men, because there was a strong legislative commitment to the abolition of sex discrimination.

Program evaluation is a more specialized type of field research, the distinction being that field research generally describes any research which is performed in a naturalistic, rather than laboratory, setting, and may include research which examines the characteristics of specified populations in the absence of specific programs, while program evaluation is limited to research which evaluates the effects of a program on specified populations. While most field research has an action component, in that it is designed to be useful to decision-makers, program evaluation usually has somewhat more focused goals regarding specific decisions about program changes or the efficacy of continuing the program under study.
Types of Program Evaluations

Perusal of the literature on program evaluation suggests that at least three distinctively different but interacting approaches may serve to describe the process and issues involved in such research. The individuals who are actively engaged in the conduct of such evaluations generally write about specific problems they have encountered, deficiencies in the field, or recommendations for upgrading and/or systematizing this relatively new area of endeavor. These writings are generally the most practical in nature of the three approaches, but often fail to adequately address the issues involved in determining appropriate methodologies and in planning for and understanding the political basis and police implications of program evaluations. The methodologist, who is generally an academician more than an evaluator, provides a more in-depth perspective upon the methodological options available for this type of research, and the shortcomings of various types of possible strategies. The methodologist, however, frequently has little awareness of the political and practical realities of conducting such research, due to their focus on technical excellence, which at times renders their recommendations applicable only to ideal state conditions. Nevertheless, this approach is valuable in that it points out the conditions which are desirable in the design of evaluations to increase the reliability and validity of the results obtained. The third approach, discussed by those individuals who are primarily interested in the policy implications of evaluation research, focuses on the political, economic, and other factors external to methodological considerations which influence the
design, conduct and outcome of such research. This approach is also largely an academic endeavor, but it is valuable in that all stages of evaluation research are highly dependent upon political processes: increased awareness of these processes will result in evaluations which are more relevant to the decisions to be made, and which ultimately have greater impact upon the policy making process.

The working definitions which are utilized by individuals from these three approaches, together with some of the major issues which they raise, are presented below.

**The Practitioner Approach:** The practitioner often gears his definitions and recommendations for improving program evaluation to specific purposes, such as increasing the level of methodological sophistication within federal agencies about appropriate evaluation strategies. Since most of the program evaluation efforts are funded by federal agencies, the bureaucrat/practitioner is a highly appropriate target group. A widely circulated manual entitled *Practical Program Evaluation for State and Local Government Officials* (Hatry, Winnie, & Fiske, 1973) provides a number of suggestions for improving the design of evaluations, as well as an overview of the current state-of-the-art in program evaluations. The authors indicate that program evaluation efforts to date have often been simply program descriptions, and that a more adequate definition of program evaluation might be "the systematic examination of specific government activities to provide information on the full range of the program's short and long term effects on citizens...while program evaluation may include considerations of workload measures, operating procedures, or staffing, its
chief focus is on measuring the program's impact or effects."

Other practitioners, speaking to a more methodologically sophisticated audience, have pointed out the disappointingly low impact of evaluation results on policy and program operations and, in response, propose the development of standard evaluation methodologies, planned linkage at the design stage between the information needed and the options the data might support, and the establishment of "cost-value" criteria for evaluation as remedies for the difficulties encountered in planning, interpreting, and utilizing such evaluations (Buchanan, Horst, & Scanlon, 1973). These recommendations imply increased staffing within (federal) departments, including planning for extensive in-house evaluation efforts rather than relying on the judgment of separate organizations contracted to perform evaluation work. Evaluation is thus viewed as a system of planning efforts, rather than unrelated single efforts. These practitioners foresee the development of an orderly, systematic process for conducting evaluation research which is not contingent upon the values of the participants, and which results may be utilized in a similarly disinterested, scientific fashion.

The Methodological Approach: The methodologist's focus is upon the methodological problems peculiar to field research and the special techniques which must be employed in such cases. This approach is even more explicit about the inadequacies of current methodologies. Campbell (1972) has noted that "with the most minor of exceptions,...none of our major ameliorative programs have had adequate evaluations" (p. 3). He further notes that "most [evaluations] are anecdotal, or document the
delivery of services without estimates of the effects of those services. The more elaborate ones are correlational [Levine, 1969]. Truly experimental ones are rare" (p. 8). By and large, proponents of this approach attempt to discuss the issues in evaluation research in an apolitical manner, as exemplified by Campbell's view of the passive technical role for the social scientist as program evaluator. "The job of the methodologist for the experimenting society is not to say what is to be done, but rather to say what has been done. The aspect of social science that is being applied is primarily its research methodology rather than its descriptive theory, with the goal of learning more than we do now from the innovations decided upon by the political process...even the conclusion drawing and the relative weighting of conflicting indicators must be left up to the political process" (p. 2).

The Policy Approach: This approach, which focuses primarily on the policy-related issues involved in evaluation research, is to some extent present and acknowledged in almost all evaluation writings. A more systematic examination of this approach, however, is provided by Carol Weiss (1973) who sharpens the focus of the political context of evaluation efforts. "The policies and programs with which evaluation deals are the creatures of political decisions. They were proposed, defined, debated, enacted, and funded through political processes, and in implementation they remain subject to pressures—both supportive and hostile—that arise out of the play of politics...since evaluation is undertaken in order to feed into decision-making, its reports enter the political arena. There evaluative evidence of program outcomes
has to compete for attention with other factors that carry weight in the political process" (p. 1).

As was evidenced by the policewomen evaluation, the goals of the evaluation, the values and priorities of decision-makers, the roles of the participants in shaping these questions, the political climate surrounding the program and the evaluation, and the current state of knowledge regarding the subject need to be considered in determining the appropriate questions.

These three directions will be discussed as they pertain to the generation, conduct and outcome of the evaluation of policewomen on patrol in Washington, D.C. in the following sections. For a more detailed discussion of the methodological and policy-related issues in program evaluation research, refer to Appendix B.

BACKGROUND OF THIS STUDY

Throughout the history of this country, women have played a marginal role in law enforcement. The first policewoman to have the power to arrest was hired in 1905, but progress was slow thereafter. In 1972, the Police Foundation estimated that only 1.5% of all police officers in the country were female. With few exceptions, the majority of these women were performing functions which were within the confines of traditional women's roles, the most common being

---

1 The history of policewomen has been adapted from Bloch, Anderson and Gervais (1973b, pp. 47-52).
clerical work, matron duty, and assignments to female and juvenile offenders. After studying the utilization of women in seven major departments across the country, Milton (1972) of the Police Foundation concluded that the major obstacle to women seeking careers in law enforcement was men's social attitudes. Virtually no women were permitted access to the mainstay of police work—uniformed patrol (the first policewoman assigned to patrol was in Indianapolis in 1968). This exclusion is particularly significant, since patrol work comprises the bulk of police operations, it is considered by many to be the most important police function, and it is also the essential prerequisite for advancement to most other positions within police departments. All men are initially assigned to patrol, where they usually spend anywhere from one to five years before competing for more specialized jobs. The female officers' lack of street experience clearly set them apart as junior officers who were ill-prepared for a wide range of duties, and tended to lock them into a narrow range of promotional opportunities.

The Metropolitan Police Department of the District of Columbia was typical in this regard. From 1918 until 1967, the Department had a Women's Bureau, which initially was founded to perform case work and (non-uniformed) patrol duty in "areas where young women might be lured into criminal or immoral conduct, e.g., railroad stations, moving picture houses, (and) theaters." The total number of policewomen rarely numbered over 30 during these years, women were required to have a college degree, and they were not held to the height requirements for men (5'7"). Policewomen did not work with male officers until 1967, when the Women's Bureau was disbanded. Most of the women in the
Department at that time were assigned to the Youth Division. In 1969, Jerry Wilson was appointed the new Chief of Police in the District of Columbia, and immediately thereafter, he initiated a program aimed at using women interchangeably with men. Recruitment campaigns were altered to appeal to both sexes, the college degree requirement for women was lowered to the high school diploma requirement in existence for men, women were required to meet the height (but not weight) standards for men, and a wider variety of assignments were opened to women, including tactical squad, investigative work, plainclothes work, and other assignments which did not require a uniform. Wilson allocated 110 positions for women, and created the job of policewomen's coordinator, which was intended to assist in implementing the program until women had become assimilated into the department. After two years, when skepticism and opposition within the department had turned to an acceptance of women in these positions, some officials even requesting women for positions formerly reserved for male officers, Chief Wilson prepared for the last step—uniformed patrol. In December of 1971, he held a press conference to announce his decision to hire 100 female officers (in addition to the previous 110), for uniformed patrol duty during the coming year. All selection, training and deployment standards were to be identical for the sexes except for weight.

One of the reasons for this move was the Chief's firm conviction that women were capable of handling the job, based on his earlier experience. Another was that the previous, more limited program still involved many inequities. Promotion was more difficult for women to obtain than men, even though the women were often better educated and
had served as much or more time on the force as the men who were promoted. Women who were able to obtain supervisory positions often suffered a lack of respect from subordinate officers, in part because of their lack of street experience. Some of the men objected to the fact that women were able to obtain certain specialized assignments, which have better hours and in some instances better pay, without their having the full street experience required for men applying to such positions. Although the policewomen program was certain to provoke opposition within the department, the Chief and the Chief of Patrol felt the resistance would be short-lived, as the women proved themselves capable.

The Police Foundation realized that the Chief's commitment to hire women provided the first opportunity to study a substantial number of policewomen on patrol, and that, if the experiment was a success, the results would be far more persuasive than Ms. Milton's journalistic investigations or the results from two extremely limited experiments with policewomen which were also funded by this agency. One of these, the first evaluation of policewomen, was performed in St. Louis, where four policewomen were placed in a high crime housing project as security guards. Although their performance was found to be more than satisfactory (Sherman, 1973), the small sample as well as the unusual duty did not permit generalizations beyond the experiment, but merely indicated that these four policewomen appeared to be able to handle some of the rougher aspects of police work. The second was roughly concurrent with the D.C. evaluation. Fourteen policewomen in New York City were reassigned to uniformed patrol and intensively studied
by the Urban Institute (Greenwald, Connolly, & Bloch, 1974). These women were observed in an extensive number of incidents, and extremely detailed information was obtained about the officers' actions which was useful in suggesting differences in styles of policing between men and women. However, the small sample size, the atypical qualifications of these women (all had been officers for a considerable length of time before going on patrol, all had volunteered for patrol duty, and all were on special hand-picked teams of [male and female] officers noted by their department for their competence), as well as the difficulty in adequately matching these women with comparable men, forbade generalizations beyond the experiment. Thus, the D.C. experiment was the first opportunity to determine whether the average woman who would be hired under normal conditions was capable of handling patrol work.

The potential impact of this study was greatly enhanced by passage of the 1972 amendment to Title VII of the Civil Rights Act of 1964, which made regulations prohibiting sex discrimination applicable to state and local agencies. The Equal Employment Opportunities Commission guidelines specify that "the principle of non-discrimination requires that individuals be considered on the basis of individual capacities and not on the basis of any characteristics generally attributed to the group" (EEOC Guidelines, 1972, Section 1604.2). The implication of these changes in federal law for police departments was that the burden of proof was now on the local department to show that sex was a bonafide occupational qualification for employment, i.e., that men are clearly better qualified for the job, and that there is available no alternative method which would allow police operations to be
effectively carried out while still permitting the hiring of qualified women.

By late 1972, it was estimated by the Police Foundation that there were about 250 policewomen on patrol in the entire country, roughly 100 of whom were on the Washington, D.C. force (including women reassigned to patrol as well as newly hired women). Wilson's plan to hire 100 new women was curtailed due to a hiring freeze, so that only 80 "new" women were available for study during the first study year. The women were assigned to two of the seven police districts in D.C. in order to facilitate the evaluation and to give the women a greater chance of equal treatment. Eighty men who graduated from the same training classes, whenever possible, and who were assigned to two other districts which were similar to the districts to which the women were assigned, comprised the comparison group. In the second year of study, six additional women who joined the department through the spring of 1973 were added to the study sample, increasing the total sample to 86 women and 86 men.

The evaluation occurred over a two year period, from 1972 to 1974, which permitted study of the entry process, the program problems, and the performance of policewomen relative to the performance of men with similar experience on the force. The first year of study presented comparisons between male and female officers who had an average of 4 1/2 months street experience. In December of 1972, the evaluators recommended that the study be continued, since this time period was insufficient to determine whether the results would hold for these officers after they had gained more experience. In addition, during
the first year of study, the officers were almost invariably patrol­ling with a more experienced partner. Extending the evaluation would permit study of these officers patrolling alone and as equal or occasion­ally even as senior partners. Chief Wilson agreed that an extension would provide valuable information, and on December 15, 1972 the Police Foundation Board voted funds with which to continue the project for another year.

INITIAL PROGRAM EXPERIENCES

In January of 1972, the first female officer was reassigned to uniformed patrol, to be followed by about five other officers during the next two months. In the spring of 1972, all of the women who had been hired since Wilson had instituted his new standards for women (1969) who were assigned to the Patrol Division (about 60 women), were slated to be reassigned to uniformed patrol. Two "retraining" classes lasting one week apiece were scheduled, to reacquaint the women with such matters as directing traffic, arrest techniques, scout car pro­cedures and report writing. At this point, there was a great deal of resistance from both the women and their commanding officers. Many of these women had jobs which were desirable because they were con­sidered interesting or had good work schedules (day work and weekends off), and reassignment meant shift work, and, for some, child care problems. Many of the commanding officers considered the women to be essential in their previous assignments and requested their retention, although few were successful. During the training sessions, some women were being reassured by their commanding officers or even the
training officers that they would never have to go on patrol, or that the assignment change would be short-lived, and they would soon return to their former positions. A number of women became quite hostile about being placed on patrol, and complained about everything from the uniforms to charging that the department was racially discriminating (the majority of the women in the Patrol Division were black). Many threatened to quit if forced to go on patrol (simultaneously, some men were threatening to quit if they were forced to patrol with a female partner). The department was steadfast, however; the women were assigned to patrol, with strong encouragement (which became virtually the rule), to remain on patrol until the end of August before seeking their old positions, and the threatened resignations did not materialize.

A number of program irregularities, or attempts to subvert the program, occurred in the first few months of 1972, including the following:

- Although women were issued both skirts and slacks, the guidelines were unclear as to when the slacks were to be worn. On a cold day, a number of women complained about wearing skirts, whereupon their sergeant removed them from street assignments to inside jobs, rather than allowing them to wear slacks.

- The first assignment given to a woman who had formerly been an undercover agent (in that same area), was a midnight shift, alone, on a foot beat (officers normally patrol with a partner for several months to acquaint them with the job of patrol, and the high vulnerability of this assignment is apparent).
• Some male partners of women were indicating to the dispatcher that they were patrolling with a woman, or that they were patrolling alone, in order to receive more backup or less dangerous calls.

• In one of the experimental districts, women were not assigned to a special "crime patrol" section because they felt women could not handle this duty.

In mid-April, just two weeks before the first (10) newly hired women completed training academy, Chief Wilson issued a 19 point order, prompted by the results of monitoring efforts by the policewomen's coordinator and the evaluation staff, which forbade any special treatment of women, either favorable or unfavorable, and specifically dealt with noted program irregularities (refer to Bloch, Anderson, & Gervais, 1973b, pp. 57-59). This effectively resolved the bulk of the problems during the first year. Between April 30 and late October of 1972, 80 newly hired women were trained and assigned to patrol with few major problems.

In August of 1972, the first charge of police brutality was leveled against a policewoman (and her male partner), both of whom were exonerated by the department. At the end of August, the policewomen's coordinator, who had been quite active in detecting and effectively taking action against program irregularities, passed the sergeant's examination and requested patrol duty. The coordinator job was then eliminated, since the major implementation problems seemed to be over. In the fall of 1972, the inspectors of the two experimental districts complained to Chief Wilson that most women were not as effective as men, and that they felt handicapped by the large
concentration of women in their districts. In response, Chief Wilson decided to conduct a survey of his own, and accommodated the evaluators by incorporating questions of interest to the evaluation, using a similar format, and by including the evaluation's matched sample of men in the survey so that there would be an adequate basis of comparison. In one of the experimental districts, some of the women lodged a formal complaint against some of the male officers, charging them with repeatedly making sexually inappropriate remarks. Their complaint resulted in animosity between the sexes in this district, and one male officer was disciplined. As the weather grew colder, the women complained about walking foot beats with summer attire. After several weeks of cold weather had passed, the winter uniforms were finally issued. In December of 1972, Chief Wilson reviewed the results of his departmental monitoring efforts, and called a press conference to announce that the policewomen program had been a success, and that women would now be hired from the same Civil Service lists in a routine fashion. Since the experimental period was over, the special orders which he had instituted during the experimental period were no longer in effect.

GOALS OF THE POLICEWOMEN EVALUATION

The major goals of this evaluation included: 1) to provide data from which to base legal arguments regarding sex discrimination within police departments; 2) to provide data to police departments, their funding agents and local officials for use in making policy decisions about the utilization of women in policing; 3) to provide data which
would be useful to the D.C. Police Department as well as other police departments in dealing with the administrative difficulties encountered in such a program; and 4) to provide an initial data base for use by other researchers with related interests.

**Legal Goals**

These were considered the most critical for the evaluation, since they carried the potential for having a national impact upon policies regarding sex discrimination (it was anticipated that few police departments would voluntarily hire substantial numbers of women for patrol duty without legal precedent or direct legal pressure). As a vehicle for enforcement of the laws banning sex discrimination, however, this study holds a peculiar position, in that: 1) performing this evaluation made it possible that, if women were found to be deficient, the results might be used to justify continuing policies of sex discrimination, removing the burden of proof from the departments; but 2) since this study was performed with a great deal of effort and thoroughness, and with some scientific credibility, it mitigated the possibility of any extremely biased or poorly executed studies performed by interested parties being presented and possibly accepted as evidence.

**Policy Goals**

Moreover, this was the first major study of its kind. It was recognized that if the findings showed no dramatic and persuasive differences between men and women, these results could be used to put pressure on police departments to either defend their employment practices or to hire policewomen. "No difference" results were likely
to increase the probability that federal and private class action suits would be levied against departments, and would provide greater reason for federal funding agencies to threaten economic sanctions if sex discrimination persisted. In addition, it was hoped that some police departments would voluntarily consider the results of the D.C. experiment in determining how best to respond to sex discrimination within their departments.

**Administrative Goals**

Results relevant to these goals may be useful in the implementation of similar programs designed to use women interchangeably with men. Nevertheless, such goals are of lesser importance because of their dependence upon the prior decision to hire women. For this study, the larger administrative aims were to delineate the major obstacles to implementation of such a program, and to suggest program management strategies for facilitating the detection of program irregularities and problems which would provide an organizational structure within which women would have a fairly equal opportunity to perform. The immediate aim was to provide the D.C. Police Department and the evaluators with ongoing information about the success with which the program was being carried out as planned, in order to minimize program deviations which threatened to jeopardize program success and/or the evaluation efforts.

**Research Goals**

Research goals were two-fold: the major purpose of conducting this evaluation with a modicum of methodological rigor was to establish the scientific credibility of this research, for enhancing the legal and persuasive goals. Secondarily, the evaluation was an attempt to
gather initial data using almost every conceivable type of criteria and measure considered as relevant, which might be useful to researchers with related interests. With respect to this second goal, the evaluation should be considered in light of the reliability and validity of the data obtained, as well as for its heuristic value. Since the study was intended to provide data about major performance differences between policemen and policewomen, and the impact of these differences on the public, a number of areas which might be interesting to the researcher were not fully explored, due to their indirect relevance to the legal and persuasive issues. For example, sex differences in styles of policing and in attitudes toward the job were only sketchily explored, since their contribution to the other goals of this study was minimal. The psychological characteristics of women entering the police field were not investigated in any direct respects, despite the frequent interest of researchers in this area, since this data was not relevant to the study purposes.

ROLES OF STUDY PARTICIPANTS

Within these broad divisions of goals, there were, of course, individual organizations and persons who were involved in the evaluation, and their roles as well as values influenced the form and outcome of the evaluation. The most salient of these were the funding organization, the evaluation contractor(s), and the D.C. Police Department. The funding agency, the Police Foundation, is a private, non-profit organization which was established by the Ford Foundation in 1970 for the purpose of supporting innovation and improvement in
policing. Their reasons for funding the study, as stated in the forward to the final report, were the following: "Because patrol operations take up the majority of the average police agency's resources and are considered the most important of all police work, the Police Foundation decided to sponsor an evaluation of the performance of women on patrol. The purpose was to provide police and public administrators with as much scientific information as possible on the subject...The Foundation believes that, while the evaluation was conducted in Washington, the results will be useful for police agencies nationwide" (Bloch & Anderson, 1974a, p. 1). This organization was a reasonable candidate for funding a study which was certain to be controversial and unpopular in many police quarters, since it is privately funded, small (having approximately 12 professional staff members), and "progressive" in orientation. The relative lack of bureaucracy and the Foundation's private status meant that this organization could act quickly in funding a study on short notice, that the research contractor would not be required to satisfy federal requirements for evaluations, which might have seriously delayed the evaluation,2 and that the Foundation would not be constrained by political considerations from seeking and presenting data which would be used as a major source of persuasion for or against the utilization of women on patrol. The study would benefit the Police Foundation in that it was likely to

2The major obstacle would have been the requirement that measurement instruments used in federally funded studies must be approved by the Office of Management and Budget prior to use. Approval often takes about 3-4 months.
receive a great deal of publicity, some of which would doubtless be negative in nature, but which would serve to enhance the national stature of this relatively new, little known organization.

The principal staff member from the Police Foundation in charge of this evaluation had a joint appointment with the Urban Institute at the time of the study's inception, which facilitated the planning stages of the evaluation. As previously mentioned, another staff member had investigated and authored a report on the status of women in policing immediately prior to the D.C. evaluation (Milton, 1972), and held a personal and professional interest in improving the status of women in policing. While others kept a more neutral stance, she was actively engaged in an advocacy position, and hoped that this study might serve to provide harder data that "women can do it" than the opinions or experience of a few mavericks in the field.

The Urban Institute, which undertook the actual evaluation, is a private, non-profit research organization which was established in 1968 to study the problems of the nation's urban communities, to respond to the needs of policy-makers for disinterested analysis and basic information, and to attempt to facilitate the application of this knowledge. Most of the initial founders had strong ties to various areas of the federal government, and this organization was founded in part to permit policy-related research to be conducted without being subject to political pressures and bureaucratic delays. The professional staff numbers about 110, a sizeable percentage of which are economists, and the staff tends to maintain close contacts with the academic as well as the federal government communities. The background, orientation
and professional contacts of the Urban Institute made it suitable as the organization selected to perform an evaluation which was expected to have a national impact, and which would need to be performed in a manner which was convincingly objective and scientific.

As an organization, the Urban Institute had no strong interest in the area of criminal justice research. Nevertheless, the policewomen evaluation was consistent with the general aims of the Urban Institute, of providing analysis of major urban problems which would assist in formulating national policies for ameliorating these conditions. Further, some of the Institute staff had been involved with research in related areas, and the principal investigator had a special interest in the reform of police services. The Institute's goals for the project therefore reflected those of the principal investigator.

The principal investigator for the policewomen evaluation had experience in and knowledge of the criminal justice field (Bloch & Ulberg, 1972; Bloch & Specht, 1974) as well as organizational contacts with the Police Foundation. The goals of this individual must be matters of inference. However, his major aim seemed to be related to developing information which might be useful in potential anti-discrimination court cases. This study was consistent with his previous interest in improving police operations through evaluation efforts designed to have a specific impact, and his training in jurisprudence. Since the results were expected to be looked at very critically when used in this fashion, great care was taken in this study so as not to bias the results in any desired direction.
In part because of the extensive publicity this study received after the first year, formal review committees, within both the Urban Institute and the Police Foundation, were formed to consult with the researchers during the second phase. Since the audience for the report varied considerably in their interests and level of methodological sophistication, one of the issues addressed by the review committees concerned the style of the report. It was agreed that the major targets for the final report were police departments and the legal profession. Therefore, the results were written in a simple, journalistic style, and were published separately from the methodology, tables and measurement instruments. The methodology was geared primarily to researchers in the field, with explanations of statistical terminology provided for the occasional reader lacking a research background.

The different priorities associated with the varying goals of this study are illustrated by disagreements during the first phase of study over whether or not to present the results of the multiple regression analysis of the Patrol Survey. The authors favored presenting the results of this analysis, even though they explained relatively little of the variance, since this was a pioneering effort in the field, and the results obtained suggested interesting, plausible relationships between attitudes and background variables. Other members of the review committee compared the regression coefficients with those obtained from other studies in more established areas, using non-attitudinal data, and found them to be disappointingly low, thereby recommending that they not be included. This example also serves to illustrate the non-statistical factors (i.e., the type of data, the
evaluation goals, and the state of knowledge of the area of investigation), which were taken into account in determining the acceptability of the statistics obtained.

In the first phase of study, the International City Manager's Association (ICMA) conducted the observations of police performance. This organization had also bid on the contract for the evaluation, and they were given a relatively small part of the study to perform, under the supervision of the Urban Institute staff. Their role, however, was considerably more important than simply to conduct this part of the evaluation, since this organization is the national organ for communication with city managers, and hence it was effective in widely circulating the study results and lending it respectability in different circles. Since it is often the case that mayors appoint police chiefs, and that they could, although they often do not, influence some of the policy decisions of the police departments in their municipalities, particularly in such matters as sex discrimination, city managers were deemed a particularly important group to reach for enhancing the impact of the study.

The D.C. Police Department initiated the policewomen program on their own initiative, without any apparent external pressure and without attempting to serve anything but their own departmental purposes (i.e., they were not attempting to serve as models for the nation, nor to provide data for litigation against recalcitrant departments). Their purposes in instigating the policewomen program are largely known to only the individuals concerned, primarily the Chief of Police and the Chief of Patrol. Interviews with these two
administrators as well as inferences from other actions taken by these individuals suggest that some of the following reasons may have been operating:

1) Both Chief Wilson and the Chief of the Patrol Division were reasonably certain that women could handle the job of patrol, due to their previous experience with using women in a more limited capacity. This program was the next logical step to take in determining whether women could be used interchangeably with men. Wilson's stated intention was to hire policemen and women from the same Civil Service lists (for the same job description).

2) The top administration appeared to be committed to reducing discrimination within the department and in other ways to developing a progressive, highly professional police department. Under Wilson's administration, a number of other "progressive" changes had occurred, such as intensifying the recruitment of black officers, with a resultant increase in their proportions within the department, expanding the roles available to women in the department, and hiring a civilian Director of Administrative Services, who was responsible for conducting the recruitment of black officers and overseeing the policewomen program (normally, police departments do not hire civilians). In other respects, such as their training in specialized areas, their weapons-control policies, and their arrest disposition monitoring system, the department was considered to be advanced in the field.

3) Enhancement of community relations by reducing sexist hiring practices, by hiring a more racially balanced police force, and by possibly defusing some of the aggressiveness and related undesirable
behaviors of the male officers which were deemed offensive to the community.

4) Upgrading the quality of the department by introducing officers who might have different styles of policing, attitudes, or general behavior. Consistent with this rationale is Chief Wilson's statement to the effect that he valued having college graduates on the force, even though they did not usually stay long, because of their positive influence on other members of the department.

5) Enlargement of the qualified applicant pool, thus permitting more local hiring and/or the hiring of higher quality applicants.

While the D.C. Police Department's reasons for conducting the program are relatively simple to infer, their goals for the evaluation of the program are not. Chief Wilson felt an internal investigation was sufficient to assess the program's outcome, but was persuaded by the Police Foundation to permit independent evaluation. Throughout the evaluation, he was very supportive of the research efforts, but he did not appear to use the evaluation for administrative or policy-making functions. His assessment of the program's success appeared to be primarily based on the departmental feedback (including the results of the Chief's Survey). For example, in December of 1972, before completion of all of the data analysis, he held a press conference to announce that the first year of the policewoman program had proven to be a success. At the close of the second phase of study, a conference on policewomen was held, to which police officials, mayors and selected city officials, legal officials and criminal justice researchers were invited. At this time, Chief Wilson again spoke very favorably about
the policewomen program, but in informal conversation indicated that he had not read the final report (although it is possible that he was aware of the findings through briefings by other members of the department). Neither did the Chief appear to be using the evaluation as a tool for nationwide implementation of the program of women on patrol. He very explicitly and strongly indicated that he did not wish to testify against other police departments, and when he was subpoenaed to testify for the prosecution in a lawsuit against the Cleveland Police Department, he very reluctantly complied, but this forced testimony threatened to endanger the good relations between the Urban Institute and the department. Nevertheless, the Chief agreed to and fully supported two years of the evaluation effort, which entailed some interference with normal routine, additional manpower expenditures to assist the evaluation effort, and violation of departmental privacy and autonomy. It is inconceivable that he was unaware of the possible uses of this evaluation, or that he would have agreed to permit an evaluation which imposed a fair number of inconveniences upon his department if he felt that it would serve no purpose, or if he was opposed to the purposes to which it would be put. Therefore, the only likely conclusion is that he felt a systematic evaluation was valuable for purposes beyond the scope of his department, i.e., that this study would address and answer a number of questions which would be useful in persuading and/or requiring, if necessary, other departments to hire women for patrol duty. His commitment to hiring women was firm enough to fully support the program and the evaluation, but he probably did not wish to personally participate in the politics or the legal battles
which would ensue from the evaluation because it might damage his position as a police administrator. Similar to other professions, there is a great deal of support within the system which precludes participation in litigation or other public "attacks" upon one's peers.

Any examination of the forces operating in such a program would be inadequate if the roles of the persons directly affected by the program were ignored. A number of analysts have written about the role of a police officer. The actual percentage of time spent in the roles of peacekeeper, crime fighter, or community service agent have been contrasted with the police officer's conception of (his) role. That is, the major portion of an officer's job is spent performing social service or order-maintenance functions (Wilson, 1968; Cruse & Rubin, 1973), yet the average officer generally characterizes his job as dangerous and exciting, with the major function being to catch criminals. Reiss (1971) notes that citizens are antagonistic towards police in only about 8% of citizen-officer encounters. Thus, a change to a more "progressive" department will have to recognize and deal with the resistance by the officer's conceptions of their role, and of perceived threats to role maintenance. It is possible that the introduction of women threatened the masculine image which is often portrayed by the media and perpetuated by the officer's self-conceptions, and that part of the resistance to women was in effect resistance to the demasculinization of their job image. This was suggested by open-ended responses to questions about the policewomen program, in which a frequent response was that police work was "a man's job."
SUMMARY

The purpose of this dissertation is to examine the process of conducting program evaluation research, using the evaluation of policewomen on patrol in Washington, D.C. as the illustrative example. Particular emphasis will be paid to the methodological issues involved in field research of this type and the policy implications of such research. This study was the first major evaluation of a sizable number of women performing traditional male functions within the police department. The conditions under which the evaluation took place were highly volatile and visible, and there was a great deal of opposition within the department to this change in role definition of a police officer. The following sections describe the specific methodological constraints, the results, and the impact of the study results to date.
METHODOLOGY

This section is intended to describe the actual process of conducting one type of field research.

The various methodological decisions which were made will be examined (including the study design and the choice of populations, measurement instruments and analytical techniques), with respect to the practical constraints which were operating, the different goals of the study, and the methodological soundness of the measures employed. It should be recognized that in any evaluation, the roles of the various participants (e.g., decision-makers, funding agents, program staff, researchers, advocates, consumers), the values and traditions of the organizations concerned, the purposes for conducting the evaluation, the prevailing political climate, the methodological strategies employed, the evaluation results, and the eventual impact of such research are intertwined, each evaluation having its own peculiar mix of factors which determines the outcome. Any discussion of the relative impact of such factors requires a certain amount of subjective judgment, and to some extent ignores the fact of their interrelatedness. Therefore, the following discussion of the factors which bore upon the methodological decision-making process should be considered to be one part of a larger network of factors which contributed to the eventual shape of this evaluation.
A variety of populations, measurement instruments and methods were employed in this evaluation in order to respond to the evaluation questions from a variety of perspectives, as noted earlier. These included different police departments who could use this data for planning their response to the policewoman question, the legal profession, which would be using this study as evidence for sex discrimination cases, the D.C. Police Department, which needed this information primarily for administrative purposes, and other researchers in related areas. The measures were differentially relevant to the varying goals and purposes of this evaluation, and the study had many of the typical challenges of field research, such as limited time, money, and populations. Unforeseen problems in administration and data collection procedures resulted in certain measures being considered as less valid indicators than others, and criteria for program success varied according to the individual or organization's interest and needs.

Due to the above-mentioned factors, it was deemed necessary to collect a wide variety of data in order to try to satisfy the different concerns of the groups involved in the project. Although some particular measures overlapped, there were basically three types of measures which were independent of one another, all of which provided an opportunity to determine whether women were incompetent -- the performance observations, the officer ratings and the interviews of citizens who had received service from these officers.
The development of and the rationale for the experimental design will be presented first, followed by the specific methods which were used to evaluate the policewoman program. The measures will be presented in the following order: sample characteristics, measures of equal opportunity to perform, performance measures, and attitude measures. Within each grouping, the methodological issues and the relevance of these measures to the goals of the evaluation will be discussed, followed by a detailed description of each of the measurement instruments and the specific problems which were encountered during the course of the policewomen evaluation with each measure. For a more detailed description of the instruments, sample specifications, and analysis of each type of measure, refer to Appendix C.

EXPERIMENTAL DESIGN

The ostensible purpose of this evaluation, as stated in the initial funding proposal, was to determine whether women were capable of handling the job of uniformed patrol. The standard of comparison for the women was the performance of policemen with similar experience on the force. The total sample of women hired from January 1, 1972 through June 1, 1973 was selected for study, and a sample of men hired during this time period was selected for comparison. Whenever possible, each woman was matched with a male officer from the same training class, in order to control for seniority. These matching procedures resulted from discussions with police personnel, who felt that the most important variable affecting an officer's skill on the
job was length of time on the force. (Some support for this notion was provided by the finding that the reassigned women, who were more experienced than the new women, consistently obtained higher performance ratings than the new women.) The first class of (10) newly hired women graduated on April 30, 1972, followed by 70 women through October, 1972, which was the latest date for inclusion in the first phase of study. Six women who were hired after this time were added to the second phase sample, yielding a total of 86 newly hired women and an equal number of men for the second year of study.\(^1\) In addition, approximately 25 women who had been on the force prior to the onset of the study and who were reassigned to uniformed patrol ("reassigned women"), were studied less intensively, since there was no comparable group of male officers who had experience on the force but no experience on uniformed patrol duty.

All of the newly hired women were initially assigned to two of the seven police districts in Washington, D.C., called the experimental (E) districts. The men selected for comparison were randomly drawn from the sample of male officers graduating during this time.

\(^1\)Three women hired during the second year of study were inexplicably assigned to duties other than uniformed patrol (tactical squad and training academy) immediately upon graduation from the training academy, and hence were unavailable for study. The remaining 86 includes some men and women who had left the force by the close of the second year of study, but who were included in appropriate data analyses. The actual sample sizes vary by year of study as well as by measurement instrument.
period who were assigned to two other police districts, designated the comparison (C) districts, which were similar to the E districts in crime rate, racial composition, and residential characteristics. It may be seen from Table 1 that the districts to which the women were assigned had a somewhat heavier workload, as measured by the number of serious offenses per officer, and that one of the E districts had a practice of one-officer patrols, whereas the other three districts had primarily two-officer patrols. The difference in patrol styles would have affected only the results of the second year's evaluation, however, since officers routinely patrol with a partner during their first few months on the force, regardless of district. Other district differences were that one of the E districts had a sizeable area of federal buildings and also a higher percentage of white residents than the other three districts (30% compared with approximately 10% in the other districts). Overall, however, these districts were quite similar with respect to the nature of patrol activities.

Demographic variables and selection criteria were recorded and analyzed by sex, although there was no attempt to match officers on these variables, since this would have effectively eliminated most of the sample, and the remaining officers would not have been representative of the sample of officers entering the department. (It should be noted that matching on the basis of selection scores might have resulted in sampling bias. See Appendix B for a discussion of this issue.) Analysis revealed no significant sex differences on the initial selection criteria (pre-employment Civil Service Examina-
Table 1

CHARACTERISTICS OF EXPERIMENTAL AND COMPARISON
POLICE DISTRICTS DURING FISCAL YEAR 1972 a

<table>
<thead>
<tr>
<th>TYPE OF MEASURE</th>
<th>EXPERIMENTAL DISTRICTS</th>
<th>COMPARISON DISTRICTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First</td>
<td>Seventh</td>
</tr>
<tr>
<td>Workload Measures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Personnel (daily average)</td>
<td>232</td>
<td>153</td>
</tr>
<tr>
<td>Robberies per officer</td>
<td>6.8</td>
<td>4.5</td>
</tr>
<tr>
<td>Serious offenses per officer b</td>
<td>50.7</td>
<td>40.9</td>
</tr>
<tr>
<td>Assignment Practices</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent of Personnel Assigned to:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foot patrol</td>
<td>12.5</td>
<td>7.2</td>
</tr>
<tr>
<td>One-officer car</td>
<td>34.0</td>
<td>3.9</td>
</tr>
<tr>
<td>Two-officer car</td>
<td>5.1</td>
<td>50.3</td>
</tr>
<tr>
<td>Casual clothes</td>
<td>14.2</td>
<td>3.3</td>
</tr>
<tr>
<td>Clerical</td>
<td>3.4</td>
<td>9.2</td>
</tr>
</tbody>
</table>

a As reported by the Metropolitan Police of the District of Columbia in its 1972 Fiscal Year Report.
b Criminal homicide, forcible rape, robbery, aggravated assault, burglary, larceny and auto theft.

...tion scores and pre-employment interview ratings), nor in their success in the training academy (police academy scores). There also were no sex differences in the number of previous jobs held or in education. The average woman was one year older than her male counterpart, was more likely to be black, and was more likely to have children.

2 The one exception to this was that the average woman took two weeks longer to pass the police academy driving test. The driving instructor felt this was attributable to the women having less experience driving prior to acceptance on the force.
This design seemed to be the best compromise available under the circumstances. The evaluation was performed at a cost the funding agency was willing to bear, and produced data bearing on sex discrimination within police departments at a time when it was politically and socially useful to have such data. A more lengthy evaluation would not have been funded at that time, and the results might have had less impact, since legislation barring sex discrimination in state and local agencies was creating political pressure to take action, even if the available data was sparse. In addition to the more general requirement that the study be performed relatively quickly and inexpensively, there were a number of other practical constraints which determined the choice of design. The two major ones are described below, together with the rationale for selecting the compromise design.

1) The policewoman program was not proposed far enough in advance of its implementation to permit pretesting of the men in the E districts prior to the introduction of women. Therefore, there was no way to fully assess the impact of the women on the performance of the men in the districts to which they were assigned. This constraint is typical of many evaluations, which are often reactive to immediate needs for information, and which thereby suffer from the lack of an adequate data base. This situation is not easily remedied, since funding for such programs and/or evaluations can rarely be counted on for extensive periods of time, due to the frequent changes in political priorities and administrations, and the political and economic costs of supporting extensive basic research
efforts. This often results in pressure for outcome data within a relatively short time span, in part so it can be used by and credited to the administration requesting the evaluation, in part because decision-makers are not always convinced (and sometimes rightly so), of the cost-effectiveness of providing a solid data base, preferring less expensive and more timely results.

2) The women could not be adequately matched with men in the same districts because the substantial numbers of women who were assigned to the E districts almost depleted the available number of personnel openings in these districts, resulting in very few men from training classes during this period being assigned to the E districts. Neither was it feasible to attempt to persuade the department to increase the personnel openings in the E districts, since this would have meant that the E districts would have been flooded with an influx of rookies needing direct supervision from more experienced officers, thereby disrupting the ability of these two districts to adequately train new officers and to effectively provide police service.

The optimal experimental design would have involved pretesting the performance of randomly selected men in the E districts (with seniority equivalent to the officers in the later study), in order to develop baseline data of the typical male officer's performance in these districts. This would be followed by later comparisons of newly matched men and women in the E districts, as well as comparisons of the pretested sample of men with the later sample of men in the E districts who were matched with the newly hired women, to
determine whether the women had made an impact upon their performance. Changes over time, however, such as changes in the pretested sample's level of competence on attitudes, or changes in selection criteria or district operating procedures might have interacted with and confounded the effect of the women upon the men in these districts. In order to determine whether such time changes had occurred, it would be necessary to include a pre and post-test sample of men in similar districts as well (such as the comparison districts), which would help account for departmental and maturational, but not changes specific to the districts over time. Such an undertaking, however, would be rather costly, as well as disruptive of the police department. Other variations of this design are possible, but all run the risk of being rejected by the department, due to the fact that the pretest data would serve little function for the department, and, even if accepted, the design might intensify the effects of retesting, including resistance to the evaluation by the officers involved.

The design which was selected, which compared men and women who were hired and trained during the same time period, but assigned to different districts, obviates the problem of the women's presence altering the "typical" male officer's behavior, but raises the possibility of district differences interacting with the results obtained. Despite the problems in interpretation posed by district differences, this nevertheless seemed more desirable for the purposes of the evaluation than the alternative, of comparing men and women in the E districts (even if there had been an adequate sample of men in the E districts). The sudden introduction of large numbers of women in
these districts created some distress, resentment, and interruption of the usual routines in these districts. Therefore, in order to compare the performance of the women with the "typical" male officer, it was necessary to select officers from districts which had not been altered by the introduction of women. This design did not allow a test of the impact of the women upon the performance of the men, but since this was not a major goal of the evaluation, this was not deemed to be a serious drawback.

During the first phase of study, the small sample of men with seniority equivalent to the women in the E districts was studied in a limited fashion. However, it was not possible to determine whether differences between E and C men were due to district differences, to sampling variations, to the impact of the women upon these men, to the supervisor's perceiving the men in the E districts in relation to the women rather than in their usual fashion, or to the interaction of some of these factors. Further study of these men would have been costly and time-consuming, and since it serve little purpose it was discontinued.

Sampling Constraints

Certain problems were encountered during the course of the evaluation which affected the validity of the entire evaluation. The most serious of these is what Campbell labels "experimental mortality - the differential loss of respondents from comparison groups." This problem will be dealt with most extensively, since it always poses a threat to evaluations which occur over any extended time period, and the likelihood for sampling loss in the groups being compared being
caused by different, rather than similar, factors cannot be adequately determined nor statistically controlled. In the policewomen evaluation, the number of officers who resigned, were terminated, or who took leave of absence due to pregnancy or military leave was relatively small and fairly equal by sex (see Table 2). Rough determinations of the relative capabilities of these officers indicated no outstanding sex differences which would have greatly affected the study.³ However, a significantly greater number of women than men were assigned to duties other than uniformed patrol by the second phase of study. In addition, there was evidence that different characteristics resulted in men and women being reassigned to inside jobs. Since the study was designed to measure the performance of women on patrol, this was relevant to this sampling problem. This particularly affected the performance observations and the survey of citizens seen by these officers.

³By October, 1973, about 3 months after data on assignments had been collected for Table 3, 11 women and 10 men (from the total sample of 86 men and women) had resigned or been terminated from the force. Analysis of these men and women revealed that the women who resigned were slightly more educated than the women remaining on the force, whereas the reverse was true for the men. However, both sexes were less likely to be married or to have children, both were below average in "written expression" on the Departmental Rating Form, and both had lower Civil Service scores, although the difference was slight with the women. Men who resigned had distinctly lower Official's Survey ratings and lower average police academy scores. Overall, these resignations may have raised the average level of competence of the remaining male officers somewhat more than for female officers. Since the samples are small, however, this effect would be slight.
### TABLE 2

ASSIGNMENTS OF FIRST YEAR STUDY SAMPLE
AFTER ONE YEAR<sup>a</sup>

<table>
<thead>
<tr>
<th>OFFICERS ON REGULAR PATROL</th>
<th>New Women</th>
<th>Comparison Men</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>In original district</td>
<td>29</td>
<td>37</td>
</tr>
<tr>
<td>Transferred to new district</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>TOTAL ON REGULAR PATROL</td>
<td>33</td>
<td>42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OFFICERS ON NEW ASSIGNMENTS</th>
<th>New Women</th>
<th>Comparison Men</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>Station duties</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>Youth</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Training Division, Communica­tions, Experienced officer's school</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Morals</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Scooter, Crime Patrol, Wagon</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>TOTAL OTHER ASSIGNMENTS</td>
<td>34</td>
<td>43</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OFFICERS RESIGNED, INJURED</th>
<th>New Women</th>
<th>Comparison Men</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>Resigned or terminated</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Injured (Administrative Duty Branch, sick leave)</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Pregnant (Administrative Duty Branch)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Military leave</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TOTAL RESIGNED, INJURED, etc.</td>
<td>12</td>
<td>15</td>
</tr>
</tbody>
</table>

| TOTAL                       | 79 | 100 | 79 | 100 |

<sup>a</sup> This information was obtained for all men and women who graduated from training academy from April 30, 1972 through October, 1972, with the exception of one woman (and the matched comparison man) assigned to a district other than the E districts. Data on assignments was obtained in mid-July, 1973, at approximately the midpoint of the performance observation period. Assignment data on the full sample of 86 men and 86 women gathered in August of 1973 are reported in the Results section. Table 2 provides an illustration of what can be expected after approximately one year's time on the force.
A second problem concerned the interaction of race and sex. Since 68% of the women in the sample were black, compared with only 42% of the men, it could not be known whether the reactions to the women were interacting with reactions to race. Direct measures of performance, such as the performance observations, would be least likely to be affected by this type of interaction, with attitude surveys and, to some extent performance ratings being the most vulnerable.

Study of this program was still considered a reasonable endeavor, however, since it was viewed as a study of the differences in performance between real male and female police applicants, under the local social, political and organizational conditions prevailing at the time and place of the study, which includes sexual and/or racial discrimination within the department and the community, as well as culturally based sex differences of the applicants, rather than a study of innate, invariant sex differences in ability to perform patrol work.

MEASUREMENT INSTRUMENTS

In order to evaluate the policewomen experiment, a variety of measures were utilized (Table 3 illustrates the measurement instruments and sample sizes in 1973). Departmental records and procedures were monitored to determine whether the men and women had an equal opportunity to perform. Men and women were compared on various performance measures, including: 1) departmental performance statistics, such as the number of arrests and traffic citations given, instances of unbecoming conduct, and number of injuries; 2) observations of incidents handled by these officers, as recorded by outside observers;
### TABLE 3
MEASUREMENT INSTRUMENTS

<table>
<thead>
<tr>
<th>NAME OF INSTRUMENT</th>
<th>SAMPLE SIZE</th>
<th>BRIEF DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Departmental Ratings</td>
<td>80 new women, 85 comparison men</td>
<td>Ratings given to police officers on the first anniversary of their appointment and included in their official personnel records.</td>
</tr>
<tr>
<td>Chief Wilson's Survey (Form W)</td>
<td>71 new women, 54 comparison men</td>
<td>Each district was asked by Chief Wilson to compile relevant statistics and to rate each officer in the sample on patrol skills. This survey was returned directly to Chief Wilson's office.</td>
</tr>
<tr>
<td>Structured Observation</td>
<td>41 tours with one woman, 12 tours with one man, 45 tours with female-male teams, 94 tours with two men, 1 tour with two women</td>
<td>Observers were hired to ride with women and comparison males and, using a standard format, note what happened, including how officers acted toward citizens and how citizens reacted to officers.</td>
</tr>
<tr>
<td>Service Survey</td>
<td>131 interviews</td>
<td>Telephone and personal interviews with people who were directly involved in an incident to which police responded. Purpose: to learn about citizen ratings of police performance and about attitudes toward policewomen.</td>
</tr>
<tr>
<td>Community Survey</td>
<td>129 resident interviews</td>
<td>Telephone interviews to determine general citizen attitudes toward policewomen.</td>
</tr>
<tr>
<td>Officials' Survey</td>
<td>59 sergeants, 25 captains and lieutenants</td>
<td>Anonymous questionnaire for patrol officials. Purposes: Obtain opinions about women and effect on supervisors of work experience with them. Obtain anonymous ratings of new women, reassigned women and comparison men.</td>
</tr>
<tr>
<td>Patrol Survey</td>
<td>15 new women, 28 comparison men, 55 other men</td>
<td>Similar to Officials' Survey, but no ratings of individual women.</td>
</tr>
<tr>
<td>Record Monitoring</td>
<td>85 new women, 82 comparison men</td>
<td>Analysis of personnel records and civil service scores.</td>
</tr>
<tr>
<td>Interviews</td>
<td>20 policemen and policewomen</td>
<td>Confidential discussions with policewomen and policemen, most of whom were interviewed twice.</td>
</tr>
</tbody>
</table>

*Includes one tour in which a comparison man worked with a reassigned woman.

3) supervisor's performance ratings of these officers; and 4) interviews of citizens who had received police service from these officers during the course of study. In addition, 5) attitudes of citizens toward
policemen and women in the districts to which these officers were assigned, and 6) attitudes of male and female officers and of police officials were obtained. (Refer to Appendix D for a sample of the types of measurement instruments used in the evaluation. For the complete measurement instruments, see Bloch & Anderson, 1974b, pp. 137-258).

The development of the measurement instruments posed some problems, since this was a new subject area, and it was not possible to rely on previously developed instruments. The actual process involved in the development of these instruments included perusal of other attitude surveys and observation forms, discussion with other researchers in the field, psychometrists, statisticians, survey researchers, computer personnel, and the like for refinement of the instruments, and incorporation of suggestions by officers and officials regarding the relevance and the appropriateness of the terminology used in the questionnaires. Scaling methods were devised which maximized consistent and ready comprehension of the response categories. Citizen interviews required brief, clear, pre-determined response categories, with the exception of a few open-ended questions. The response format used on the officer rating forms and the two police attitude surveys consisted of lines which were unmarked except for the midpoint and the extremes. Respondents were to place a mark on the line at the point which they felt best described their feelings. Marks were later transposed by means of clear templates to numerical values. Within the three officer rating forms, the same verbal labels (e.g., unsatisfactory, average and excellent) were used for each item to be rated. This
scaling procedure had the advantages of being easily readable, of minimizing differential interpretations of the response categories, and of permitting respondents to more freely indicate their choices, minimizing verbal, numerical or other position response sets. In addition, it permitted comparisons between the abilities being rated. The disadvantages of this method are that it is more time-consuming to prepare the data for analysis than if the scaled responses were numbered in advance, it does not allow the instrument to be directly keypunched, which introduces coding errors, and it is more difficult to interpret the results to the intended audience than if each response had a specified verbal label.

Given the extremely tight time constraints under which this study was conducted, it was not possible to obtain data regarding the reliability and validity of the various instruments utilized. However, since the instruments were developed using fairly standard procedures, with attention to clear, simple scaling methods, the questions were generally straightforward, requiring no further interpretation, and the populations to whom these measures were administered were sub-samples of the actual populations about which inferences were to be made, indices of reliability and validity were not as necessary as in laboratory research. The lack of instrument validation is typical of program evaluation efforts, since there is generally little time nor money available for such purposes. This is not entirely due to poor planning, since the need for independent criteria is not as acute when, as in most program evaluation research, few theoretical inferences are to be drawn from the data regarding underlying traits, patterns, etc.,
which the questions are assumed to be measuring.

In this particular study, all of the measures except the Departmental Rating Form were administered at least twice. The fact that the results were similar over these two time periods suggests that the measurement instruments were reliable. Comparisons of similar phenomena obtained through different measurement instruments provided some evidence of concurrent validity. For example, ratings of the same officers, using different measurement instruments, different performance dimensions, and to some extent different raters, were positively correlated, suggesting some consistency in the manner in which individual officers were perceived. The results from the performance observations (1973) and citizen's perceptions of these incidents also were similar.

A practical, rather than statistical, approach to monitoring the quality of measurement instruments is provided by the example of the requirement for all studies funded by federal agencies. Rather than requiring statistical instrument validation, all instruments are reviewed by the Office of Management and Budget for clarity, relevance to the issues addressed by the evaluation, and lack of overlap with previously performed studies. This approach, if capably administered, is in many ways more appropriate to field research efforts, the major drawback being that political sensitivities of the time might have more influence in the initial planning stages of evaluations than if statistical criteria were the standard of acceptability.
In any evaluation which involves comparisons of the performance of two or more groups, it is essential to determine whether the groups being compared had an equal opportunity to perform. This was particularly important in the policewomen evaluation, since the operations of police departments are sufficiently flexible to permit extensive non-random variations in assignment and opportunity to handle various types of calls. It was known from the beginning of the program that a significant number of persons within the department were strongly opposed to the use of women on patrol, and that special care was needed to ensure that women received equal treatment. Therefore, the program was monitored very closely in the initial stages. This included: 1) monitoring of daily assignments at the district level during the first few months of the program, to determine and correct program irregularities before the evaluation was well underway; 2) supervisor's recording of individual officer's daily assignments over a six month period in both study periods. The types of assignments listed included scout car patrol, tactical squad and station assignments, as well as whether the officer was patrolling alone or with a partner. Specialized types of assignments such as juvenile work, detective work and morals were combined under "other-specify," since the study was initially designed as a one year effort, and it was not anticipated that these assignments would occur with any great frequency. While this information was valuable for noting assignment changes over time, it did not cover the period during which the
performance observations occurred in 1973 (the summer months). Practical difficulties in scheduling women for observers to ride with during the second evaluation year led the researchers to the realization that assignment to the Patrol Division did not necessarily mean that the officer was regularly performing patrol duties. Information about formal transfers (to other districts or divisions) was known; however, an officer could be also reassigned within the district to certain other duties, such as to station or juvenile work (as distinguished from the Juvenile Division), while still remaining in the Patrol Division. In addition, an officer might have a regular patrol assignment, such as scout car patrol or a foot beat, but be given daily assignments to the station or other duties as needed. 3) Hence, in July of 1973, primarily for the purpose of facilitating the structured observations, the researchers requested information from each district about the current regular assignment of all officers in the sample. This information was updated in August, 1973, and information on resignations and terminations was obtained again as late as October, 1973. 4) Roughly one month after this determination, Chief Wilson gathered evidence from his department regarding the 'formal' assignments (e.g., to the Patrol Division) of the officers in the study in preparation for a court appearance in which he was subpoenaed as a witness for the plaintiffs. 5) During the first phase of study, the (dispatched) calls assigned to the study sample were recorded over a two month period. Recording was primarily for the purpose of selecting incidents which these officers had handled which would be suitable for interviewing the citizen(s) involved, but also served as a check on
dispatching procedures. 6) The types of incidents in which observers were present were compared by sex of officer, to determine whether men and women were handling similar types of dispatched incidents. 7) Interviews with officers were conducted in part to monitor program irregularities.

The major outcome of the monitoring efforts was that most program irregularities were detected early in the first year prior to the new women being assigned to the districts. Few assignment differences occurred during 1972, the only notable difference being that slightly more women than men were assigned to station duty. During the second year, however, when the program was no longer an experiment, and Chief Wilson's special orders regarding equal treatment were no longer in effect, differences in assignment became quite marked, with women increasingly being placed in the station rather than on uniformed patrol. Thus, after the average woman had been on the force for about one year, traditional sex role distinctions were reappearing. This trend is important in that it highlights a number of problems common to program evaluation efforts, including the following:

1) The previously mentioned problem of differential attrition of comparison groups. The fact that fewer women were actually on patrol than men poses particular difficulties for the performance measures, since they were based on a biased sample of officers remaining on patrol.

2) The general tendency of programs to produce effects during the experimental or demonstration stage which disappear as the program becomes a routine operation (Suchman, 1967).
3) The tendency for programs to be subverted, given sufficient opposition, if monitoring and enforcement activities are discontinued.

4) The issue of group differences in attitudes, roles, aptitudes, etc. prior to the onset of a program exerting influence on the outcome of a program, quite apart from the treatment variables. In this instance, there were probably a number of such differences which influenced the program, most particularly the male officer's opposition to any change in women's roles within the department. The different life experiences, personal expectations, and the different roles which the men and women had partaken of prior to becoming an officer (including the maternal role for women, over half of them having children), probably influenced their attitudes and behavior on the job to some extent.

PERFORMANCE

There were several methods of assessing officer performance, including: 1) departmental statistics on arrests, disposition of arrests, traffic citations, letters of commendation and complaint, etc., 2) performance ratings by the officers' supervisors, 3) independent observation of officer performance in actual patrol situations, and 4) interviews of citizens who had received police service from these officers.

Each of these methods has particular drawbacks, which will be discussed when presenting the methods, as well as reasons for being included in the study. One major problem in performance evaluations, however, the problem of defining appropriate criteria, overrides the
problems specific to the methods selected. This is particularly true in the area of criminal justice (e.g., Logan [1972] concluded, after reviewing 100 studies purporting to evaluate the effectiveness of various programs for the prevention or control of criminal or delinquent behavior, that only 9 had evaluated a well-defined technique or program, only 3 of these 9 used proper control groups, and only one of the remaining 3 had a measurable definition of program "success").

One of the most serious obstacles to defining criteria for "good" and "poor" police performance is that there is no consensus among experts in police and criminal justice circles about such criteria. The criteria which are commonly used tend to be poorly defined and multiple in nature. This is partly due to the fact that a large variety of complex skills are necessary to effectively perform the job of a police officer, which may range from skill in dealing with people to the ability to restrain and disarm violent persons, to knowledge of the legal, civil, departmental and social service procedures which apply for different situations. In addition, the job requirements permit a great deal of discretion in methods of handling issues (Wilson, 1968), without the concomitant education and apprenticeship usually found in other "professions." (The period for training police officers is only from 3-6 months, which is not sufficient to be highly skilled in all of these areas.) While there have been some attempts to quantify officer performance into the various sub-steps required for each type of activity (Charney, 1972) most police personnel would agree that there is no single best method which can be delineated in advance for most situations.
Even if performance standards could be completely preordained and quantified, the problem of determining priorities or "weighting" performance factors remains. For example, a study of officers in 3 urban police departments found that officers failed to make a felony arrest in 43% of the observed incidents where there was probable cause (Maltz, 1972). The reason suggested for this was that the officers may have felt that the "overall benefit" of the community would best be served by not making an arrest. Since police rely on community support to carry out their work and to legitimize their authority, they cannot afford to alienate the community. The relative desirability of different styles of policing, as exemplified by such techniques as making arrests vs. calming the situation, will often vary depending upon the goals and values of the audience (the community, the department, the courts, the individuals involved, the evaluators).

The laws themselves are generally reflections of the mores of the larger community. Thus, one of the functions of police departments is to enforce the formalized aspects of the prevailing morality of the populace. In addition, police officers are a product of this society, since they, more than most "professionals", interact with all levels of society. Police departments as a whole vary in their values, partly due to the composition of the community and of the members of the force, and partly due to the value positions, and ability to enforce such positions, of the top management within the department. Due to the close interface between the police and the community, the multiplicity of functions of police officers and their virtual monopoly on force, they are affected by societal values and at the same time are a
potentially powerful force in facilitating or hindering societal change (Silverman & Kim, 1972).

The criterion problem is further accentuated by the fact that each type of performance measure has limited applicability as an indicator of performance. The major liabilities of each of the types of measures used in the study are indicated below: departmental statistics generally indicate very little about performance, since their purpose is primarily administrative; officer ratings are prone to personal bias and halo effects, and usually have rating categories which are so vague that the link to performance is unclear; direct observation of performance, generally considered the most valid of such measures, must by its nature involve observer effects; and consumer surveys indicate only what consumers feel they know and what they are willing to publicly acknowledge. Telephone and personal interviews are restricted in the number and complexity of the questions which can be asked.

In order to meet some of the objections arising from each type of performance measurement, performance was assessed from a variety of perspectives, with a variety of measurement instruments, using a variety of unweighted indicators of performance. It was recognized that none of these performance criteria were sufficient in themselves, but believed that this approach would provide decision-makers with a wide variety of evidence with which to make policy decisions, based on the relative importance they ascribed to the various indicators within their own system of goals and values.
PERFORMANCE RATINGS

A common means of evaluating program effectiveness is through ratings of the performance of those included in the program. The most common type of performance rating is that used in employment situations, in which the immediate supervisor completes a standardized employee evaluation form. This form usually contains judgments of the quality of the employee's work, which is then used as part or all of the basis for promotion, salary increases, disciplinary action, termination, and the like. While employees may participate in this process to some extent, the supervisor is ultimately responsible for the ratings. Other types of performance ratings are conducted primarily, although not exclusively, for research purposes, and generally have no impact upon the person being evaluated or upon the evaluator. These include ratings which are completed by supervisors for explicitly experimental rather than organizational purposes, peer ratings, self-ratings, and outside "expert" ratings. The extent of accountability for these ratings, the anonymous, interactive or contractual nature of the rating process, and the relevance of the ratings for the persons involved all affect the outcome of the ratings. In addition to these factors, the content of the rating categories per se, which varies from fairly abstract theoretical constructs to specific behaviors, influences the reliability, validity and ultimate uses of the ratings.

Supervisory performance ratings have certain unique characteristics which are not present in other types of performance measures, and which suggest their use as a performance indicator. They are the only
type of measure which provides an evaluation of overall performance ability, rather than time and/or situation specific behaviors. They permit an "informed" opinion from the persons who are most familiar with both the behaviors required for performance of the job and with the performance of the persons being rated. They involve a minimum of time and expense compared with other types of performance measures. Because of their wide use across a variety of situations, ratings are readily understood by a general audience. In the evaluation under study, it was particularly important that the results be credible and comprehensible to police departments, who might view unfamiliar measurement instruments, criteria, and statistics with suspicion, but who would be more likely to accept and understand the general import of the results from rating forms. While ratings are obviously insufficient as the sole criteria for performance evaluations, they nevertheless provide a valuable, albeit imperfect, contribution.

Performance ratings suffer from a number of problems, however, which limit their usefulness, including the following:

1) Ratings generally lack operationally defined criteria for the activities being measured, and many rating scales do not even specify behavioral attributes. This results in high vulnerability to differential interpretation of the meaning of the rating categories, halo effects, and/or outright personal biases influencing the rating process. Although there have been recommendations for behaviorally defined scales in performance appraisal (Epstein & Laymon, 1973), this generally has not been implemented.
2) Many rating forms lack normative guidelines regarding the percentage of individuals who should fall into each category, which permits wide variance between supervisors' standards for acceptable performance. Even when rating forms have such guidelines, they are often not adhered to.

3) Ratings which occur in different locations, districts, schools, etc., may not be comparable if different informal norms are operating in the different locations.

4) When ratings are obtained for experimental purposes, and these ratings must be performed by persons within the organization, department, etc., rather than by trained outside observers, the ratings obtained will depend on such factors as the organizational support for the evaluation and/or project, the incentives and disincentives for completing the ratings, the relevance of the questions being asked, the stated purpose of the evaluation, and the rater's purposes in completing the evaluation form.

5) If all aspects of the rating process are not supervised by the evaluators, irregularities can easily occur, such as persons other than those designated to perform the ratings or to be rated being included in the sample, and completion rates will suffer, in some instances being difficult to accurately ascertain.

Most of these problems were noted in at least one of the three surveys containing officer ratings. As previously mentioned, the men and women under study were assigned to different districts, which could have affected the ratings if district norms for performance ratings differed. However, the similarities between the districts'
operational procedures, the strictly defined organizational structure of police departments, the crossover in personnel between districts, and the similarities in district characteristics (crime rate, racial composition and residential characteristics) tends to mitigate district differences significantly affecting officer ratings.

Officer ratings were obtained from the Departmental Rating Form, the Chief's Survey, and the Official's Survey. All of these measures were given to officials who had some familiarity with the officers in question. However, there were some essential differences, as described below.

**Departmental Ratings**

The Departmental Rating Form, which was obtained from the officer's personnel jacket, is the standard departmental form which is completed by an officer's immediate supervisor at the end of the officer's probationary year. Therefore, results from this form were available only during the second phase of study. This instrument has the advantages of being familiar to the officials, of being restricted to ratings of officers who have been on the force long enough for officials to have an informed opinion and for the officers to have stabilized their behavior, and of the ratings having real consequences for the officers, forming part of the basis for salary determination and career advancement, for which officials are accountable. The categories in this rating scale, however, suffer more than the other two rating forms from the problem of extremely vague performance categories, such as "learning ability" and "knowledge and skill," which require a great deal of inference from observable behavior and which have an uncertain
relationship to actual performance on the job.

**Chief's Survey**

The Chief's Survey was instigated by Chief Wilson for the purpose of conducting an internal evaluation of the policewomen's program, following complaints by the district commanders in the E districts. The Urban Institute assisted in constructing the format and content of the survey so that the results might be useful to both parties, with a minimum of duplication of information. This survey requested a listing of the officer's daily assignments over a six month period, as well as other statistics, such as the number of arrests and letters of commendation or complaint. Unlike the Departmental Rating Form, this survey was designed for experimental purposes, and the officials performing the ratings were not required to identify themselves. The format was unfamiliar, and the ratings had no job-related implications for the officers or officials involved. The fact that it was distributed by and to be returned to the Chief of Police, however, lent it credibility within the department and probably enhanced completion rates, as compared with rating surveys which were to be returned to the research organization. (In the first phase of study, 93% of the Chief's Survey ratings requested for the newly hired women and comparison men were returned. In contrast, only 73% of the ratings requested on the (anonymous) Official's Survey, which was to be mailed directly to the Urban Institute, were completed and returned.) Despite the Chief's support for this survey, however, the fact that the survey was part of the policewomen evaluation rather than serving a stable departmental function, and the lack of accountability for the ratings,
may have resulted in officials considering this survey less seriously than the departmental form. This was suggested by the fact that several officers were asked by their supervisors to complete this form themselves. This problem could have been circumvented had the researchers been able to directly supervise the administration of these questionnaires. However, it was not possible to gather a sizeable number of officials together at one time to complete the ratings. Police personnel rotate between shifts every two weeks, have different days off, are assigned to different districts and assignments within the districts, and, in the summer months when the data was being collected for the second study period, vacation leave was frequent. This left the alternative of providing each district with clear guidelines as to the administration of the surveys.

The researchers only became aware of the self-ratings during the second phase of study, but several officers confirmed that they had been asked to complete their own rating forms during the first phase as well. This posed serious validity problems for using data from this instrument, which were not dispelled by the rather elaborate mechanisms developed for detection of the self-ratings. During the second phase

---

4Ratings of officers who had indicated (in officer interviews or on the survey), that they had completed the ratings themselves were deleted. When two or more surveys from the same district had handwriting that appeared to be identical, they were presumed to have been completed by the same official and were considered valid (several surveys indicated the name of the official who had completed the form, and a comparison of handwriting from surveys with the same official's name verified that one individual had completed the forms). A number of surveys had handwriting dissimilar to any other surveys from the same district. Several of these surveys also had the officer's name written on the top of the survey. In marginal cases, when it was still unclear, the handwriting from the survey was compared with the handwriting of the officer, which was generally available in his/her personnel jacket.
of study, suspect forms were returned to the districts with reiterated instructions that they were to be completed by an official. Since this produced no rating changes, suspected self-ratings were analyzed separately. Analysis of the suspect ratings revealed that officers were more generous in rating themselves than officials had been with the remaining sample. All suspect ratings were significantly higher than the non-suspect ratings, and suspect ratings of male officers were significantly higher than for female officers. Despite the extensive precautions taken to ensure that only ratings completed by officials were used in the analysis, it could not be absolutely determined that all self-ratings had been deleted, nor whether any ratings were inappropriately deleted. Even in the event that the researchers correctly determined the self-ratings, it could not be known whether the samples of men and women who were asked to perform self-ratings (and hence were deleted from the sample), were similarly distributed with respect to the total samples from which they were drawn. Thus, the results from the remaining samples may not be representative of the total sample.

Another problem which occurred with this survey concerned the content of the rating categories, which were much more specifically tied to performance than the categories on the other rating scales (e.g., "handling a disorderly male"). This posed some difficulties for officials who had not had sufficient direct contact with the officer to observe them performing such behaviors. In some cases, this was noted on the questionnaire, and difficult ratings had not been completed. This was primarily true for officers who were no longer on patrol,
however, which supports the assumption that the rating categories on this scale were more operationally tied to performance than those on the other rating scales, for which ratings were rarely incomplete. In considering ratings of officers who were still performing patrol duties, this specificity should be considered an advantage rather than a liability.

**Official's Survey**

The Official's Survey contained both an attitude survey and a rating form. This rating form was anonymous, and was to be mailed directly to the Urban Institute, rather than channelled through the department. This form was briefer than the other two, with only four rating categories. It was to be completed by all officials who were familiar with the officer, rather than just the officer's immediate supervisor, which permitted a broader range of opinion regarding an individual officer. The format and procedures utilized in these ratings obviated some of the problems encountered in other rating forms, such as officers rating themselves, officials rating persons on abilities of which they had little or no knowledge, and officers' ratings being solely dependent upon the opinion of one official.

Discussions about the effects of repeated measures tend to focus on the validity of the measures obtained the second time of testing, due to the subjects being influenced in their responses by the initial testing. The Official's Survey, as well as others, illustrates another major effect of retesting—a decline in willingness to cooperate with the evaluation. This was particularly acute when there was no departmental directive to complete the survey, and when the researchers were
unable to oversee the administration of the survey. The motivations of the individuals concerned in completing the surveys probably played a part, in that 1) previous officer ratings and more direct expressions about the policewomen program had not served to alter the Chief's course. Prior to the second evaluation phase, the Chief had announced the policewomen program to be a success and had indicated his intentions of continuing the program, which may have been sufficient disincentive for expressing opinions about the program and/or its participants. 2) Most of these same officials had completed this same survey previously, and it is likely that they had participated in other ways in the evaluation, including completing other surveys, which may have caused them to become weary of the evaluation. 3) Specific differences in attitudes of district commanders and/or officials toward the evaluation itself may have been instrumental in determining whether or not the surveys were completed, as suggested by the fact that return rates on a number of surveys differed widely and consistently by district, particularly in the second study phase.

This survey also exemplifies some of the advantages and disadvantages of anonymous ratings. Lack of accountability tends to lower the return rate, but may elicit more genuine expressions of attitudes. At the same time, it is not necessarily the best method for eliciting "objective" ratings which could be backed up by evidence, if need be. The finding that the women were rated least favorably on this anonymous survey, most favorably on the Departmental Rating Form, for which officials were clearly accountable, and mid-way between these two ratings on the Chief's Survey, for which accountability
was somewhat ambiguous, is consistent with the notion that attitudes, but not necessarily objective reports, are more readily elicited by anonymous surveys. The Chief's Survey was distributed by district commanders, to be completed upon the Chief's directive and returned to his office, which may have led some officials to believe that their ratings would be known, rather than anonymous. (A few officials indicated their names on this survey, even though this was not requested.) Thus, the Official's Survey may have served as a vehicle for expressing opposition to the program, rather than being viewed simply as an assessment of the officer's capabilities.

Departmental statistics on a variety of performance measures were obtained from the officer's personnel jacket, including the number of job-related traffic accidents, injuries and days of light duty since appointment, the number of felony and misdemeanor arrests, recorded instances of unbecoming conduct, and the number of complaints and commendations received from citizens. Data was obtained from the Chief's Survey on a number of similar items: the number of felony and misdemeanor arrests and moving traffic violations recorded in the district to which they were assigned, favorable and unfavorable letters or comments from the public, and, in the second phase of study, the number of service-related disabilities, the number of days spent on light duty or absent from work due to service-related disabilities, and the number of injuries suffered by the officer's partner and by other officers who were present with the officer.
Duplication of information from these two sources was in part accidental, the information requested on the Chief's Survey having been instigated by Chief Wilson's desire to assess the ongoing progress of the program, and in part deliberate, since the Chief's Survey did not cover the entire period during which the women were on patrol. This duplication ultimately proved to be useful in acquiring a fuller picture of these performance variables, since each source had specific limitations which were not fully known in advance, rendering one source, or the combination of information obtained from both sources, more valid than the other.

The statistics obtained from these sources seem to be objective and bias-free performance indicators at first glance; a closer examination reveals that even these measures were subject to difficulties in measurement and interpretation. The special problems encountered with some of the departmental statistics are indicated below.

The problems in interpretation which can result from judgments based on seemingly objective measures are illustrated by the findings regarding "letters of commendation or complaint." Letters from the community and the department are routinely recorded in the officer's personnel jacket, and were also requested on the Chief's Survey, the latter broadening the definition to include letters or comments. During the first period of study (as of October 31, 1972), the women appeared to have an astounding number of unfavorable "letters or comments" from the community, according to the Chief's Survey results (63 compared to 1 for the male sample). Closer investigation revealed that the vast majority were comments, as recalled by the women's
sergeants, rather than letters, and the researchers suspected that
sergeants were expressing their own views more than those of the
community. In support of this assumption, the personnel jackets, which
were examined shortly before the Chief's Survey was recorded, indicated
no written complaints for either the women or the men as of that time.
In contrast, women had received more written commendations than the men
(7 to 3), according to the personnel jackets, and an identical number
of favorable comments or written commendations (18), according to the
Chief's Survey. Despite sampling differences (personnel jackets were
obtained for 58 men and 58 women, whereas this information was recorded
on the Chief's Survey for 78 women and 69 men during the first study
phase), the discrepancy in unfavorable letters or comments from the
two sources is so large as to preclude accepting the Chief's Survey
data.

Statistics regarding service-connected injuries and days absent
from work or on light duty (due to service-related disabilities) suf­
fered from other problems in interpretation, that of possible differen­
tial treatment by sex for injuries and of differential opportunity for
sustaining such injuries. Informally, there were known instances of
women with minor injuries (such as alleged back strain because of a
gun belt being too heavy) being placed on light duty when men with a
similar level of disability probably would not have requested or been
granted light duty. Thus, days of light duty were not necessarily a
reliable indicator of severity of injury. Further, since the women
were not placed on patrol duty as frequently as the men, they had less
opportunity to be injured, on the average. To circumvent this problem,
the rough criteria of having spent no more than 30% of the time on 
inside assignments (in the six month period measured by the Chief's 
Survey) was used as the definition for officers who had primarily been 
engaged in patrol duties. Officers on patrol were then compared by 
sex. Somewhat greater accuracy would have been obtained from comparisons 
utilizing the actual total days on patrol per officer, although this 
approximation was probably sufficient for the purposes of this evalua-
tion.

In the first phase of study, the average male officer felt that a 
female partner would increase the risk for her male partner more than a 
male officer with similar experience. Consequently, in the second 
phase, the total number of injuries received by partners of the officers 
under study was added to the Chief's Survey. This measure proved 
fruitless, however, since this information is not normally recorded in 
departmental records, and hence a mechanism for collecting this infor-
mation would have had to have been in effect at the onset of the study. 
Other sources for this information included the Structured Observation 
instrument, to which injuries to the officer or his/her partner were 
added in the second phase, and the investigation of major incidents 
by the evaluators. Since both of these sources covered only a very 
limited portion of the officer's performance on the job, and since 
officer injuries are a relatively rare event, virtually no information 
about this question was obtained from these efforts.

The number of arrests made by an officer were very questionable 
indicators of performance, for the following reasons:
1) In order to make comparisons between groups, the groups must have an equal opportunity to make arrests. Since the women did not patrol as frequently, they had less opportunity to make such arrests. The rough division of patrol versus inside assignments was made from the Chief's Survey, and officers whose primary assignment was patrol work were compared. As discussed before, this is only a rough approximation, and other factors, such as district differences, the specific beats assigned, the calls dispatched, the proportion of foot beats (which carry less opportunity for arrests) to patrol car assignments, determine the true "opportunity" for arrests. As a whole, the women were patrolling in districts which had a slightly higher crime rate than the male's districts, although one of these districts had a sizeable area consisting of government offices, in which there is relatively little activity. Recording of dispatched calls by observers indicated that the women received more calls, and that the calls were of at least as serious a nature as those received by the men, but there were some observed instances in which dangerous calls were diverted from female officers. Overall, it could not be determined with precision whether the women had an equal opportunity to make arrests, although a rough weighing of the factors involved suggests that the women who were regularly assigned to patrol assignments probably did.

2) Police department reporting procedures permit only one officer to be credited for each arrest. While some partners are rather egalitarian about the distribution of arrests, others, particularly if they are the senior partner, will take credit, since some feel it is advantageous to have arrests on one's record. Given the antagonism
towards the women, and the more passive role that they sometimes played, it would not be surprising if women were not fairly credited with their share of arrests. Measurement of this requires direct observation, however, and there were too few arrests made during the times when observers were present to test this hypothesis.

3) Accurate arrest and arrest outcome data were difficult to obtain for a variety of reasons, some of which were attributable to the organizations involved having no efficient, functioning procedures for retrieving this information for evaluation purposes, others of which were more politically based. There were three sources for arrest data, including: 1) The Chief's Survey, which requested information about the total number of arrests made by the officer since assignment to the district; 2) the district arrest books, from which trained "statistical clerks" recorded all arrests credited to the officers under study since assignment to the district as well as the disposition of the arrests; and 3) computerized data from the Prosecutor's Management Information system ("PROMIS") on the type and outcome of all arrests for serious offenses made by the officers under study over the period of a year's time (Alprin, 1973).

The Chief's Survey data is less complete than the data obtained from the other two sources, arrest information having been completed for only 81% of the female and 61% of the male officers. Nevertheless, the proportion of felony and misdemeanor arrests made by male and female officers for whom this survey was completed were similar to the proportions indicated from the other two sources.
Arrest data from the departmental arrestbooks were initially used to determine whether the disposition of arrests made by men and women was similar. At this time when this information was recorded, it was uncertain whether the PROMIS data would be available. The recording procedures employed by the statistical clerks involved selecting the arrests made by the officers under study from the total sample of arrests made by all officers in that district (ordered by date of arrest, rather than by officer). This method proved to be highly tedious and error prone, most notably in errors of omission, and consequently, all district records were recorded twice, by two different clerks, to achieve a higher rate of accuracy. Clerical errors were compounded by omissions in the arrest books (5% of the arrests listed had no disposition listed, and another 10% had incomplete dispositions. Although the results of this method were reported, the PROMIS data were considered more accurate and complete.

The PROMIS system, which was a recently constructed, computerized system, developed for the use of the police department and the courts for improving feedback about arrests, and well publicized as an advance in the criminal justice field, was extraordinarily difficult to implement. Despite the fact that the PROMIS data was hypothetically available and the system functioning, it took several months of negotiating to obtain this data. Numerous delays were encountered in obtaining clearance from the prosecutor's office to gain access to this data system, in part from bureaucratic inertia, and in part because of prosecutor's sensitivity about the political implications of making data about their activities public (when clearance was finally obtained, the
prosecutor's office requested a review of any material stemming from the use of these tapes prior to publication). Once clearance was obtained, it was discovered that the output was unuseable, requiring negotiations with programmers within the police department to correct the program errors so that officers could be plotted with their arrests and dispositions. It was apparent that this seemingly functioning system was not currently in use, since these programming errors would have been discovered. When correct outcome data was finally obtained, the discrepancies between the PROMIS data and the arrest book data were considerable, and could not be reconciled by the missing outcome data from the latter source.

4) Arrests per se cannot be said to be desirable, unless it can be shown that they were both necessary and legal. As previously mentioned, arrests made by the officers under study were examined to determine whether the type or outcome of arrests differed by sex of officer. Data tapes were obtained from the PROMIS system about all arrests for serious offenses made by the men and women during the year beginning August 1, 1972.

5 Offenses for which the maximum penalty is a fine of $300 or less are not included in the PROMIS system, which primarily excludes arrests for disorderly conduct and for municipal violations, such as permitting a dog to run without a leash. Offenses which had been dropped by the officer before reaching the prosecutor are not included.

6 The first new women went on uniformed patrol on April 29, 1972. Computer data on arrests made before August, 1972 were considered inaccurate and were not included on the sample. Arrests which lacked an arrest date but which listed an offense date later than August 1, 1972 were included in the analysis.
felony or misdemeanor arrest was made, the most serious offense as well as the most serious disposition for any of the listed charges were recorded. Men and women were compared on their conviction rate (a disposition of "guilty"), as well as on the other types of arrest outcomes. The reasons for "no-papers" (the least serious disposition, in that the case is rejected outright by the prosecutor), were also recorded and analyzed by sex. The assumption underlying these recording procedures was that, if an arrest was considered legal and the evidence sufficient, the case would be considered suitable for prosecution and would not be dismissed by the prosecutor or the judge (no-papered, nolle, or dismissed), prior to trial. "Good" arrests were considered those in which a conviction occurred, even though the conviction may have been for a lesser offense, which was generally the case. This procedure provides an approximation of the legality of an arrest, rather than an in-depth study of this issue.

The question of whether an arrest was necessary, however, is not readily amenable to evaluation efforts, since it involves a combination of knowledge of the law and knowledge of the context of the situation, including the psychological state of the participants in the incident.

---

7 Crimes against persons were considered the most serious type of offense; narcotics offenses were considered more serious than crimes against property.

8 Dispositions were ordered in terms of seriousness by considering how far the offense had been carried in the judicial process. "Guilty" was considered the most serious disposition, followed by "not guilty," "dismissal" (by the judge after the prosecutor had entered the case), "nolle prosse" (the prosecutor initially accepted the case and later rejected it), and "no-paper" (initial rejection by the prosecutor).
Arrests may be provoked by the officer's behavior or avoided by skillful intervention. Officers may choose to ignore arrests in many situations when they would be warranted. The measurement of the actual necessity for an arrest is extremely difficult, is not readily repeatable by other researchers, and is time consuming, since arrests are not a frequent occurrence. Suggestions regarding possible ways to measure stylistic differences in handling provocative situations are provided by the performance observation instrument used in the New York City evaluation (Greenwald, Connolly & Bloch, 1974), but it should be noted that the analysis of such detailed data proved to be cumbersome and expensive.

DEPARTMENTAL PERFORMANCE-RELATED STATISTICS

In assessing the usefulness of departmental statistics which bear upon performance, the following issues in measurement should be considered:

1) The accuracy of recording and measurement procedures.

2) Whether the data recorded covered a similar time period for the groups being compared.

3) The extent to which personal biases could have affected the recording process.

4) Whether the type of measure is a reliable and valid indicator of performance ability.

5) The effect of extra-experimental variables upon the statistics obtained.
6) Whether groups had a differential opportunity to receive such measures.

7) Whether the measurement process was impeded by other factors, such as the measurements themselves requiring a high degree of sophistication and judgment or the events being measured occurring with relative infrequency.

Despite the potential problems with statistics which are obtained from departmental sources, these measures are of value since they are the most readily available performance measures, they have been used in countless other evaluations for various purposes, and many of these statistics are likely to be available on a continuing basis, well after the completion of the evaluation. This provides an opportunity for followup that can be readily implemented with a minimum of expense and expertise in the conduct of evaluations.

PERFORMANCE OBSERVATIONS

Performance observations are probably the most valuable, valid and reliable indicators of performance. The most apparent advantages of this type of measure are that 1) they are direct measures of actual performance on the job, in contrast to performance ratings, consumer satisfaction surveys, and job simulations (such as training academy scores), which require a fair amount of inference from the measures to later performance. 2) They sample a wide range of the typical behaviors required on the job, rather than focusing on discrete performance indicators, which are subject to variable interpretation. 3) Performance observations are one of the few types of measures which
can, if designed properly, take into account the interaction between performance skills and the context within which this performance takes place, including the opportunities for performance and the interaction between the individuals under study and the individuals who they may be dealing with in the course of their work. 4) Observers, who generally have little personal stake in the outcome of their observations, can be trained to be fairly consistent and unbiased in their observations, whereas the typical respondent in other types of measures is an employer, consumer, administrator, or other individual who may have some personal interest in the survey results, who is usually not trained or otherwise constrained to be "objective." 5) Another obvious advantage is simply an artifact of the measurement instruments utilized in performance observations. If observation instruments are properly formulated and observers well-trained, a great deal of detailed information can be recorded in a relatively brief time period, whereas survey formats, wherein individual questions must be asked, considered and responded to (by respondents who are generally untrained), are much more time-consuming, and therefore permit a more limited number of questions.

In summary, performance observations permit measurement of relatively complex behavioral indices of performance, with knowledge of the context in which these indices were obtained, and with relatively little personal judgment and involvement to distort the data recording process.

There are certain problems with performance observations, however, which limit their usefulness as the sole means of assessing performance.
Although a wide variety of behaviors are recorded in performance observations, there is generally little opportunity for "summing up" or weighting the performance indicators by persons knowledgeable about a more extensive period of performance and about the behaviors which are most important on the job. Thus, they may be insufficient as the sole measure of overall performance, even though their contribution is invaluable.

Performance observations often pose many more administrative difficulties than other measures, and they are very time consuming and expensive to conduct. Recruiting, selecting and training observers, scheduling observations so as to sample representative situations but to minimize program disruption, conducting extensive observations and supervising observers is considerably more difficult than administering pre-coded surveys to selected populations. Depending upon the purpose of the evaluation, it may or may not be cost-effective to obtain such detailed information.

In most situations outside of laboratory research, it is not possible to eliminate the possibility of observer presence affecting behavior, nor to measure this effect. Despite the relative objectivity of outside observers, there is still the possibility of observer bias due to the cultural, political, occupational or other characteristics of observers. This can be minimized in a variety of ways, such as 1) constructing the observation instrument so that subjective biases are less likely to occur, with the use of behaviorally based rather than inferential items, 2) selecting observers so as to minimize and/or counterbalance likely biases in the area of study, 3) training
observers to attain consistency in recording, and 4) utilizing statistical checks to determine whether observer characteristics had an influence on the outcome of the observations.

- Certain types of data may be important but difficult to obtain through performance observations, since they necessarily sample a limited time period. For example, in evaluations of controversial programs, the performance of typical participants in typical situations may be only part of the information which is necessary for making policy decisions. Exceptional performances, or performance in critical situations, may be as important to program outcome as performance in the average situation. Critical situations, however, may occur so infrequently that an adequate sample of performance in such situations will be difficult to obtain.

The problems noted above are illustrated by the experiences in conducting the performance observations of policewomen and their male counterparts. A description of the observation program, followed by the specific methods chosen to achieve some degree of resolution of the problems inherent to performance observations, will be presented. Since this was considered to be the most important measure of performance, the procedures involved will be described in greater detail than for the other measurement instruments.

Trained observers rode in scout cars with the matched male and female officers under study for a total of 191 evening shift tours. (Each tour is eight hours. Evening shifts included the 3-11 p.m. and the 4 p.m. to midnight shifts.) For each tour, observers recorded demographic data on the officer and his/her partner, if present, such
as age, date of appointment to the force, and the number of times the (male) officer had patrolled with a policewoman, together with data about several aspects of how the officers worked together, if patrolling with a partner. Separate forms were completed for each incident in which some significant interaction occurred between officers and citizens, which included detailed descriptions of the characteristics of the incident and the citizens involved, the actions of the officer under study and his/her partner, and each citizen's reaction to specific officer actions. Relatively brief incidents in which no significant officer-citizen interaction occurred (consisting primarily of false alarms, incidents which were handled primarily by other officers on the scene, and officer-initiated incidents such as traffic stops, bank and business checks, chats with citizens, etc., in which citizen reactions were neutral), were recorded on a brief incident form, which noted only the type of incident and the major action taken by the officer(s). Observers completed the Community Survey before their first ride and again upon completion of their last ride to determine whether initial attitudes toward policewomen had affected their reporting, and to see if attitude changes had occurred as a result of their observations. Upon completion of the last observation tour, observers were interviewed or completed an interview form (Ride-Along Summary), to supplement the data available on the structured observation instrument.

Observations as indicators of overall performance: The tendency for performance observations to provide a great deal of data, but no priorities, was in most respects advantageous. This permitted
different police departments, which have different goals and values and which serve different populations, to draw their own conclusions about the data, based on their needs or on the criteria which they consider most crucial to their operations. In addition, the "objectivity" of the evaluation might have been compromised by advocacy of particular criteria; in order to render the evaluation suitable for use by the courts, it was advisable to present all of the data without assigning priorities. Finally, the fact that summary measures of overall performance ability were available from other measurement sources provided alternate indices which could be compared with the observational data.

**Administrative difficulties:** Administrative difficulties were numerous, some of the more salient being that:

1) In the first phase of study, the selection, training, scheduling and supervision of observers was conducted by the International City Management Association (ICMA) rather than by the Urban Institute, which served in a consulting and supervisory capacity but did not directly handle the daily administrative problems. As a consequence, certain decisions were made regarding the design and conduct of the observations which differed from the original design set forth by the Urban Institute staff. This arrangement can be viewed as a trade-off between political impact and methodological desireability, the ICMA being influential in increasing the impact of the study results, but direct supervision being more desireable for coordination of data collection efforts. By the second phase, the political impact had been assured, and at that time the Urban Institute was given full control of the observation process.
2) The original samples of 86 men and 86 women were differentially available for observation. During the second observation period, approximately 48 women and 62 men were on patrol assignments, and of this number, only 33 women and 48 men were actually observed at least once during the two month observation period. Men were observed only in the C districts, but due to the extreme difficulty in observing women, observers were sent to all districts in which women were on patrol.

In order to succeed in scheduling observers with the designated officers, extensive preparation was necessary, including: a) a liaison was established with an official in the Chief's office to provide official sanction for the observations and to assist in resolving problems which occurred in the districts (e.g., a special directive to permit female observers to ride alone with male officers, which is normally against departmental policy, was necessary for the observations); b) at least one official at the level of lieutenant or captain in each district was designated as the official contact in the district, to which requests for the particular officers to be observed, requests for information on officer's assignments, and complaints about daily problems encountered in scheduling observers were conveyed; c) lists of all of the officers eligible for observation were given to the districts, with instruction that the officer's current assignment, shift and days off were to be designated. From these lists, which were frequently updated through observer and district feedback (including information about vacations, injuries, assignment changes, etc.), observers were given lists of the names of officers who should be
available for observation, together with instructions on random selection procedures; d) districts were contacted daily to determine whether the pre-designated officers would be available for observation, and observers were to contact the Urban Institute prior to each observation tour to obtain updated assignment information; 3) The ratio of observers to officers who were expected to be available was generally 1:3 or 2:5-6 per district, to ensure that one of the designated officers would be available; f) In the event that no officers were available, even after these precautions, observers were given alternate districts in which to observe officers. Generally this meant that observers scheduled to observe women were rescheduled to the C districts; g) After initial difficulties in scheduling observers with policewomen became apparent, attempts to balance the number of tours with male and female officers were instituted, at times even initially scheduling twice as many tours with women as with the men.

Despite these precautions, more tours with male than female officers were observed (55% compared with 45%), and 11% of all attempted observations were unsuccessful. Because the number of women on patrol was less than the number of men on patrol, the average woman was observed more frequently than her male counterpart (2.6 times compared with 2.2 times). The sample of women under observation probably had less patrol experience, on the average, than the sample of men under observation, since in a number of known instances, women who would have normally been assigned to the station, a foot beat or the wagon were assigned to a scout car to accommodate the evaluation, whereas no special attempts were made to increase the number of men available in the scout
cars, and 2) results from other surveys suggested that the males on patrol may have been slightly more competent than the average male officer, whereas the reverse was true for female officers. The difficulty in scheduling observers with women was due in part to the fact that women who were assigned to the patrol division were often given daily assignments in the station, the paddywagon, or on foot beats, which were not observed due to the differences in the nature of the work tasks and the obtrusiveness of the observation. The relative infrequency with which women were assigned to scout cars seemed to be due in part to the fact that no women had a permanent scout car, whereas a number of the comparison men had permanent cars. (A few women were relief officers, which meant that they patrolled in a scout car when the permanent officer was sick, on vacation, or had days off.)

3) Certain attempts to change the usual police department routines to accommodate the observations, even when sanctioned by the Chief of Police, were unsuccessful. For example, during the second phase of study, most of the officers were sufficiently experienced to be certified to patrol alone. Although there should have been equal proportions of male and female observers by sex and type of patrol car, previous departmental policies to "protect" female observers (or male officers) prevailed, in that female observers were never placed in scout cars having one male officer (only 31% of the policewomen patrolling alone were observed by women). Thus, departmental tradition inadvertently resulted in an interaction between the characteristics of observers and the individuals under study.
In some cases, overt resistance to the observation program was encountered. When this was suspected, it was virtually impossible to schedule observers on that day, since all observations were performed on evening tours (during the second phase), and the liaison officials were generally off-duty by the time that observers were certain that they would not be riding with any of the officers in that district. Disruption of police department routines from the observation program was heightened during the second phase because of the extensive publicity surrounding the D.C. experiment, which resulted in a large number of persons from other police departments, journalists, researchers, students and others requesting rides with the female officers. Eventually, the frequent observation of the women in the program led to some overt anger on the part of both officials and women in the program towards the observers.

4) Hiring and scheduling police observers proved to be far more difficult than with civilian observers, since this job represented "double duty" and did not pay commensurately, because officers' shifts changed every two weeks, posing scheduling problems, and because replacement of police officers was considerably more difficult than with civilian observers, the former necessitating another departmental solicitation.

**Observer presence:** Although there was no certain way to eliminate the effects of observer presence on performance, there were a variety of attempts to minimize this effect, including: training the observers to be as neutral as possible, to avoid conflict with the officers over such matters as whether the incident was sufficiently safe for them to
be present, and to avoid interacting with officers and citizens during incidents. Given these precautions, the most likely effect of observer presence is that officers might have refrained from expressing socially undesirable attitudes or behavior. Citizens were probably minimally affected, as suggested by the fact that observers frequently reported being mistaken by citizens to be police officers. It is not known, however, whether sex differences exist in the reaction to observation, or if the fact that the average woman was observed more frequently than the average man may have resulted in different observer effects. A comment by one of the female officers who was interviewed during the program illustrates this problem (Bloch & Anderson, 1974a). "The women in the Department are over-spotlighted. If we could just be left alone, we would perform better. It's always 'Who's looking at us now? Who's in the paper now?' That kind of thing really wears you down after awhile" (p. 49).

Observer bias: In order to minimize observer bias due to sex, race, occupation, or personal/political orientation, the following measures were instituted: 1) observers received extensive training with a standard set of stimuli (films of police incidents, training manuals, and the Structured Observation instruments) to increase the accuracy of their perceptions, to familiarize them with police procedures, terminology and typical situations, and to produce a high level of consistency in their understanding and recording of police events. 2) An attempt was made to hire equal numbers of male, female, black and white observers. During the second phase of study, police officers as well as civilians were hired so as to provide a broader perspective upon
police performance, to lend greater credibility to the results, and to enhance the training and experience of civilian observers. During the training period, police observers were more accurate in recalling the details of a situation, and they were less likely to term actions as hostile or disrespectful than civilian observers. These differences seemed to be largely due to their experience on the force—police observers used their prior experience with citizen reactions to police officers as a baseline (i.e., the "average" citizen reaction from their experience was labelled neutral), whereas civilian observers seemed to be using different norms, such as the reactions of a friend or coworker. For ratings which required some subjective judgment, training sessions focused on discussions of the characteristics which could be consensually agreed upon, such as displays of anger, friendliness, disrespect, etc. The general rule in making these decisions was to minimize inference on the part of the observer, relying on behavioral indicators of such emotions rather than on personal interpretations of how the individual must have felt in the particular situation. Individuals selected to be observers were briefly screened with respect to their feelings about policewomen, and those having strong biases for or against policewomen were not hired. Observers were rotated between districts, and attempts were made to balance the number of male and female officers observed by each observer. The observation data was analyzed by race and sex of observer and officer to determine whether these demographic characteristics had influenced the reporting process. The relative absence of differences resulted in the discontinuance of plans to analyze observations by observer's attitudes, since the sample of observers was
small and it seemed likely that this type of analysis would yield little additional information.

**Critical questions:** The most serious doubts raised about women on patrol concerned their ability to handle situations which involved danger and which might require the use of physical strength, weapons, or courage. The stereotypical female qualities of passivity, fearfulness and physical weakness were brought up as barriers to performance in these situations. There were fewer reservations about the ability of women to handle many of the more common occurrences in the life of a patrol officer, such as writing reports, talking with citizens and handling traffic accidents, and hence their ability to protect themselves, citizens and other police officers loomed as an important issue. After the first evaluation phase, this question was still unanswered, due to the low frequency of such events. Therefore, in the second phase of study, attempts were made to maximize the expected frequency of such occurrences being observed by conducting observations only on evening shifts, which have the highest rate of activity, by eliminating the day with the lowest activity rate (Sunday), and by emphasizing weekend tours, which have the highest probability of violent or potentially violent activity (observers were not accepted if they were unable to observe on weekends, and a slight monetary incentive was added for weekend observations). For similar reasons, the recording process was more extensive for incidents which might bear on this question than for incidents of a relatively minor nature.
CITIZEN SATISFACTION WITH SERVICE

During the first year of study, an attempt was made to circumvent some of the possible biasing effects of observer presence, as well as to obtain another measure of officer performance from a group which is often neglected, but which should be considered in determining criteria for performance of any type of service—the consumers of such services. Individuals who had contacted the police, and whose call had been answered by the male or female officers under study, were interviewed about their encounter with these officers (Service Survey). An attempt was made to question these individuals on matters similar to those observed during the tour, such as hostility-friendliness, use of force, how well the officers worked together as a team (if appropriate), and their feelings about the quality of service received. However, the interview format required considerable changes. Most notably, it was necessary to design an interview which was relatively brief and which did not request highly personal information in order to elicit cooperation. Questions had to be phrased simply and clearly so that respondents, who lacked the training of the observers, would be likely to have a consistent understanding of the intent of the questions. The total number of questions had to be narrowed considerably to ensure cooperation from the individuals being interviewed. Questions which might have been viewed suspiciously or as an invasion of privacy had to be avoided, thus curtailing information about the citizen's role in the incident as well as information about personal matters such as income.
A number of explanations and questions were required to orient the respondent to the purpose of the study, to establish the legitimacy of the interviewer, to elicit recollection of the incident, to determine whether respondents correctly recalled the officer(s) under study, and to obtain cooperation for the interview, which further abbreviated the number of pertinent questions which could be asked.

The administration of the citizen interviews posed many problems, the most serious being the difficulty in obtaining a sizeable sample of persons to be interviewed who had been involved in a non-trivial incident, and who recalled the incident and the officers with sufficient accuracy to be satisfied with the reliability of their response. Given the circumstances (i.e., that the citizens may have been upset at the time of the incident, that the incident had occurred roughly one month prior to the interview and that officers tend to appear alike because of their standard uniform), it was anticipated that citizens might have some difficulty in differentiating the officers under study from other officers who might have arrived on the scene. It was not anticipated, however, that respondents would be unable to correctly identify and recall the incident, the interview was conducted. Respondents were asked to identify the race and sex of the officer and his/her partner, if applicable. Police department records indicating the officers who had been assigned the call were matched with other records indicating the race and sex of these officers; these records were crossed against respondents' recollections. If only one officer was present (according to departmental records), and the respondent correctly identified with officer's race and sex, or if the officer was patrolling with a partner and the sex of both officers and the race of at least one officer was identified, with no incorrect identification of race (don't know was permissible), the criteria for "correct identification" of the officer(s) were met.
identify the race and sex of both officers even when no additional officers arrived. During the first evaluation phase, when the men and women under study were patrolling with partners in virtually all of the incidents sampled (99%), incorrect officer identification occurred so frequently (about half the time), and seemed to have so little relationship to whether additional officers arrived, that the researchers questioned whether the officers assigned to the call had in fact been the officers on the scene. It was suspected that police department records were unreliable for determining the officers who had been present at an incident (one officer may have remained in the car, other officers may have arrived on the scene earlier, or handled the situation, etc.), since respondents would be likely to remember events which had occurred within the month. Furthermore, even if respondents had correctly identified the officer(s) by race and sex, this provided no guarantee that these were the officers under study. Results of this survey were analyzed separately for those who had "correctly" and "incorrectly" identified the officer(s) under study, with virtually no differences between the groups. Therefore, in the second year of study, only incidents in which the observer, rather than police records, noted that the officer was indeed present were selected for interviews, in order to be certain that respondents were in fact talking about the officers under study. The possible effects of observer bias were

---

10 When two officers were present (the officer under study and his/her partner), respondents "correctly" identified officers 50% of the time. When 3 officers were present, the officers under study were identified 57% of the time. Thus, additional officers seemed to have little effect on ease of officer identification.
deemed far less serious than the problem of interviewing citizens about the wrong officers. This procedure improved the record of "correct" recall of officers patrolling with partners to 76%, and of officers patrolling alone to over 90%, which lent credence to the suspicion that police records were an unreliable indicator of the officers who actually handled the call. Unfortunately, the procedure utilized in the second year of study had a major unforeseen drawback—the sample of persons from which to elicit interviews was severely attenuated, which resulted in 131 completed interviews during the second phase, compared with 507 during the first (refer to Appendix C for a more complete description of this sampling problem).

During the second year, only data from incidents in which officers were correctly identified were reported in the major findings. In these incidents, 64% of the female and 16% of the male officers were patrolling alone. Thus, the results from this instrument tend to be a measure of citizen satisfaction with two male officers vs. one female officer. This represents a possible source of bias, since partners of these officers were often more experienced, and males tended to be patrolling with partners more often than females.

ATTITUDES

The attitude surveys used in the policewomen evaluation were not directly relevant to the legal goals of producing information relevant to the issue of sex discrimination, since the preferences of employers, coworkers or consumers (citizens) are not grounds for exemption from sex discrimination legislation (EEOC Guidelines, 1972). As research
data, attitude surveys are notoriously fraught with difficulties—the link between attitudes and behavior is variable, the construction of the questions asked greatly affects the responses received, and results are subject to varying interpretations. Nevertheless, surveys of citizens and police were useful in assessing public and departmental response to a controversial change in hiring, and of determining whether these attitudes changed over time as policewomen became less of a novelty, and as experience with these officers increased.

The major goal of these surveys was to determine the effect, rather than the effectiveness of policewomen, and as such the primary purpose was to improve program administration. In addition, it was recognized that the program was being conducted in a social/psychological context, and that to focus solely on performance variables would provide only a very partial picture of the program, divorced from the human elements which affected the participants and the community. In order to even determine whether the women were given a fair opportunity to perform, it is necessary to know more than their assignment patterns. If expectations from the department and the community are sufficiently negative, this surely poses difficulties for the women in being treated equally. Thus, indicators of the psychological conditions under which the women were functioning are valuable in assessing their performance. The attitudes of the policewomen themselves toward the job and about the performance capabilities of women vis-a-vis men were of lesser importance, but in some cases helped to explain some of the differences in performance, and in others served as a check to balance the opinions of the male officers.
Police Attitudes: Two very similar attitude surveys were used to determine police attitudes—the Official's Survey, which was administered to sergeants, lieutenants and captains in the E and C districts, and the Patrol Survey, which was administered to male and female officers on patrol assignments in these districts. Questions were asked about the performance ability and traits of male and female officers in various patrol situations, as well as about attitudes towards policewomen and the policewoman's program, job satisfaction, community cooperation, and officer-supervisor relationships. Pre-test measures of men's attitudes were obtained, although they were inadequate for anything but rough comparisons because of the small sample size, the mixture of officers and officials sampled for this survey, the higher mean age of the pre-test sample, and the different form of the instrument itself. (The typical pre-test question had verbal labels for all possible responses, whereas the questions on the Patrol and Official's surveys were marked only on the midpoint and extremes; in some instances, the pre-test had different questions, or questions which were phrased somewhat differently, than the later Patrol and Official's surveys.)

The major comparisons were between male and female officers and between male officers and officials. Male officers in the E and C districts were also compared. However, the comparison districts, which were intended to be districts in which the attitudes of male officers would be "uncontaminated" by experience with women on patrol, in fact had small numbers of women by the second study phase (probably not exceeding 4 per C district), and since they were highly visible,
they presumably could have influenced attitudes somewhat. Officers in the E districts, who were presumed knowledgeable about policewomen, had not necessarily worked with women, although most had probably seen women in action. Thus, there was no clear division of "experienced" and "inexperienced" male officers. Even if opinions were to be solicited solely from officers who had worked with policewomen in the E districts, and from officers having no work experience or direct contact with women in the C districts, these might have been biased samples, since assignment to work with policewomen did not appear to be completely random. In some instances, it appeared that policemen with strong negative feelings could avoid assignment to policewomen, which may be good program management but which contaminates the experimental conditions to some extent.

Despite the differences in measurement instruments, samples, and district, the responses by male officers and officials on the Patrol Survey, the Official's Survey, and the pre-test (on similar questions), were remarkably similar, which probably attests to the officer's strong feelings about policewomen, regardless of the district to which they were assigned or the format of the question.

Attempts were made to determine the background characteristics which might have affected attitudes, including rank, age, race, educational level and sex. Certain of these factors may have been confounded, however. For example, within the sample of male patrol officers, age may have been interacting with education and race, in that older officers who had not been promoted above the officer level are unlikely to have attained more than a high school education, and, because of
earlier hiring practices, are more likely to be white. In addition, the fact that the majority of the women were black, whereas the majority of men were white, makes it difficult to disentangle attitudes based on sex alone from racial attitudes.

Attitudes of both male and female officers were also assessed by selected interviews of officers at various times during the program. These interviews were intended to provide a fuller picture of the program and of the officers, but were not amenable to statistical analysis.

Community attitudes: The effect of introducing policewomen on the community was measured by assessing community attitudes in the E and C districts before program implementation, at two later time periods during the first year of study, and at the end of the second period of study. The attitudes of the business community in these districts were also assessed, using the same measurement instrument, during the same three time periods of the first year. The Business Survey was discontinued in the second year, since residents' and business persons' attitudes were remarkably similar. Approximately equal proportions of businesses in the E and C districts which would be considered high, medium and low crime risks were surveyed, to determine whether the "objective" risk influenced attitudes. A total of 513 community residents and 120 business persons were interviewed over the two years of the study.

Citizens were asked about the safety of their neighborhood, their general attitudes towards police, and their degree of contact with policewomen (e.g., knew a policewoman personally or had observed a policewoman handling an incident, had seen policewoman on patrol, and/or had heard about policewoman through the media). Respondents were asked
whether they felt policewomen or policemen would have an advantage in handling a number of patrol situations, and what the effect on crime and criminal apprehension would be if half the force were women. Finally, respondents were asked whether they would prefer an all-male or a mixed sex team, if they needed police service, and their attitudes toward equal rights and hiring women on an equal basis with men for police work.

Interviews were conducted by Chilton Research Services, a professional survey research firm, whose extensive experience was helpful in anticipating the problems which would be encountered in attempting to conduct an unbiased survey with a sample representative of the community. The training and interview process was supervised by the Urban Institute staff. Two such problems will be described in greater detail.

In order to sample a sizeable number of persons over several time periods, interviews were conducted by telephone, which is considerably less costly than personal interviews. This procedure biased the sample somewhat, since persons lacking telephones were not eligible. In order to avoid further biasing the sample by including only persons with listed numbers (unlisted numbers being fairly common in the District), respondents were selected by a computer-generated random number telephone sampling procedure, which gave every telephone in the selected exchanges (exchanges falling primarily within the E and C districts) an equal probability of selection. Since some of the exchanges overlapped slightly with areas outside of these districts, addresses of respondents within these exchanges had to be compared with street maps
subsequent to the interview (25 of the 154 interviews completed in 1973 were eliminated due to this factor).

Population data regarding the racial and age distributions for these districts were obtained in order to define sample specifications for a representative sample of community residents. Since the proportions of white respondents were too low to permit reliable racial comparisons (about 14% of the population in the E and C districts are white), approximately twice this number were sampled.

The procedures which were followed to ensure that the age/sex/race quotas were filled in an unbiased manner included: 1) systematically rotating interviewers among the different age/sex/race groups; 2) limiting daytime calling to avoid oversampling older persons; and 3) pre-determining the age and sex of the person to be interviewed, and if the person was not at home, making up to 4 calls back, rather than selecting an individual who was available at the time, to avoid a bias towards selecting persons who stayed at home.

These procedures, including the oversampling of white respondents and the random number selection procedure, resulted in a sizeable number of fruitless calls. For example, in 1973, only 11% of the numbers dialed resulted in useable, completed interviews.

RELATIONSHIP BETWEEN BACKGROUND AND PERFORMANCE

A great deal of detailed data was obtained about the individual officers under study which was initially utilized only for group comparisons. In the second phase of study, an extensive attempt to determine the relationship between background, selection and performance
variables was undertaken. First, background and performance variables on the Service Survey were factor analyzed; this resulted in little information except that a number of performance variables were related. Following this effort, data which were available about the individual officers from all of the measurement sources were entered into a computerized file labelled the "Super-Personnel File." The types of variables selected included background characteristics, selection and training scores, departmental statistics (e.g., the number of accidents, injuries, arrests, days' sick leave, favorable or unfavorable letters, instances of unbecoming conduct, and assignments), officer ratings from all three sources, and selected performance variables from the performance observations and from interviews of citizens who had received police service from these officers. As a preliminary measure, separate correlational analyses were performed on background (including selection) and performance variables. When a number of variables with similar content were found to be highly correlated, the variable which most generally represented this content area was selected for subsequent correlational analysis between the background and performance variables. Analysis of the relationship between performance and background variables was performed separately by sex. Comparisons of officers by assignment, within sex, were performed, using chi-square tests of distribution differences.

In order to provide a rough picture of the typical male and female officers who were considered "good" or "poor" by the department, the three best and worse officers for their sex on three criteria were compared, for a total of 18 officers of each sex. The criteria were
the Departmental Rating Form "overall" ratings, the Official's Survey rating of general competence on patrol, and the number of misdemeanor arrests. These officers were then compared on a number of other background and performance variables, in order to provide a descriptive picture which was valuable primarily for heuristic purposes, since statistical comparisons were inappropriate with this sample size.

SUMMARY

A number of methodological problems have been discussed in this section, many of which the evaluators were aware of at the beginning of the study, others which arose during the course of the evaluation. It should be recognized that any study, including those performed in the laboratory, can be dissected in a similar fashion. Field research tends to be more vulnerable to methodological criticism than laboratory research, but on the other hand it has the advantages of being carried out in real situations, with subsamples of the real population to whom the results will later be applied. Inevitably, the judgment as to the value of this type of research, in spite of the problems, is a subjective one. In the policewoman evaluation, if the primary objective is kept in mind (i.e., to determine whether or not the evidence indicated that women were not capable of handling the job of patrol), the methodology must be considered adequate to the task. A variety of performance measures were used to address a variety of goals, from the perspectives of the D.C. Police Department, the community, and trained outside observers. (It should be noted that Campbell [1970] suggests using multiple measures when multiple objectives are to be
addressed.) Despite the problems of individual measures, and the shortcomings of any one perspective, had there been any major evidence of women's inability to perform patrol, this would probably have been discovered.
In order to understand and interpret the results from this evaluation, it is important to delineate the different questions which followed in attempting to respond to the major study goals.

The major objectives of this evaluation, and the roles of study participants, have been discussed in detail in the Introduction. Briefly, they may be summarized as:

1) legal goals — to provide information useful for litigation regarding sex discrimination within police departments;

2) policy goals — to provide data which would give police departments throughout the country an unbiased picture of how women perform as patrol officers, so that departments might consider how best to utilize women in their particular situation;

3) administrative goals — to provide the D.C. Police Department with ongoing information about program problems, and to assist other departments which wish to integrate women in a similar fashion; and

4) research goals — to present initial data, together with measurement instruments and some possible approaches for evaluating police performance which may be useful to researchers having similar interests.
This section will be relatively brief, since it is not the purpose of this dissertation to report the evaluation findings per se. However, since the results of the study were critical to the evaluation impact, some understanding of the major findings is necessary in order to discuss the implications of this type of research and the evaluation impact. For a more complete description of the results, refer to the Final Report (Bloch and Anderson, 1974a), the first year's report (Bloch et al., 1973b), and the data tables from each of these years (Bloch and Anderson, 1974b, pp. 49-136; Bloch and Anderson, 1973, pp. 48-100).

METHODS OF DATA ANALYSIS

A variety of detailed background and performance data about the officers under study, as well as attitudinal data regarding community and departmental reactions to policewomen, were obtained during the course of the evaluation. The most important comparisons for the determination of future departmental policies as well as for determining whether sex was a bona fide occupational qualification were the comparisons of the performance of newly hired women and comparison men. Since sizable numbers of women were no longer assigned to patrol duty by the second phase of study, comparisons by type of assignment were often necessary to ensure that the samples had similar opportunities to perform. To determine whether initial performance and attitude data changed as the officers gained experience and the program lost its novelty, comparisons were made
between the results from the two evaluation phases. Other types of analyses were performed primarily to determine whether factors other than sex of the officer had affected performance or attitude results. The supervisory ratings given to reassigned women and the newly hired (male and female) officers under study were also compared.

During the first phase of study, t-tests of mean differences were utilized to determine whether group differences would have occurred by chance \( (p < .05) \). After review of the data, it was decided to use chi-square tests of distribution differences during the second phase, to minimize the influence of extreme scores. Trends were reported \( (p < .10) \) for important comparisons when sample sizes were small.

EVALUATION QUESTIONS AND FINDINGS

This evaluation was designed in part to respond to the legal issue of whether sex discrimination was warranted by the requirements of the job. In order to address the legal issues, it was necessary to determine whether the employment conditions were similar for the sexes. That is:

Did the women and men have the same opportunity to perform?

If selection or training criteria differed by sex, or if assignments, calls and other working conditions varied by sex, then the performance findings would be biased towards the group having the more favorable entry characteristics or work situation. Legally, it is not clear
whether evidence of harassment or strongly negative attitudes would be admissible evidence of women having less than an equal opportunity to perform. If so, then the attitudes of coworkers and supervisors, and possibly of the community, would be considered to address the legal questions regarding equal opportunity.

**Equal opportunity:** Selection criteria included a Civil Service examination and a pre-employment interview, neither of which showed any sex differences. Similarly, there were no sex differences in police academy training scores, nor in educational level. This suggests that the entry level competence of the men and women under study did not differ in any major measurable respects. The problem with these measures, however, is that previous studies have shown little relationship between selection criteria, training academy scores and performance on the job (an excellent critique of the major studies in this area is provided by Kent and Eisenberg, 1972).

Typical of the findings in this area is the seven year study of the New York City Police Department, in which one of the major findings was that performance on the Civil Service examinations predicted too little but performance on later Civil Service promotional examinations (Cohen & Chaiken, 1973). However, this study used only the performance statistics which were available in the department for performance criteria.

In the policewoman study, the relationship between the selection and training criteria utilized in the D.C. Department and later performance on the job was assessed, with the following results.

1) The interview rating was a poor predictor of performance --
"general suitability" did not correlate with any background or performance variables for the women, and correlated only with "poor" performance for the men (low average ratings on the Official's Survey and poor reactions from victims). 2) Civil Service test scores were positively correlated with education, police academy scores and race, but had few correlations with performance (women with high Civil Service scores received higher departmental ratings on writing ability, were less likely to use sick leave, and passed the police academy driving test more quickly, whereas men with high scores were more likely to have received a written commendation from citizens, were less likely to have missed work due to service-related injuries, and were less likely to have resigned from the department).

3) Average police academy scores fared somewhat better. For both sexes, academy scores were positively correlated with Civil Service scores, education, and the overall departmental rating. Men with higher scores made more misdemeanor arrests, and women with higher scores passed the driving test more readily.

Sex differences in time spent on patrol assignments would necessarily indicate differences in opportunity to perform specific patrol skills. In addition, since work experience is generally considered to be an important factor in the attainment of job skills, the degree to which men and women shared the same experience on the job was important to assess. (Reassigned women received significantly higher ratings than the newly hired women on both the Chief's and Official's Surveys in both phases of study, which probably reflects their having been on the force for a longer period of time.)
During the first phase of study, after the average woman had been on the force for four and one-half months, assignments differed little by sex (women were assigned to the station 10.7% of the time, which was about twice as often as men). By the second phase, however, women were on patrol considerably less often than the men, and were assigned to the station or other inside assignments with much greater frequency. By August, 1973, only 45% of the women remained on regular uniformed patrol, compared to 71% of the men. Inside assignments (station, youth division and community relations) were held by 31% of the women, but only 12% of the men. "Other street assignments," including morals, tactical squad and scooter squad, were held by 12% of the women and 4% of the men.

The sample of women who were placed on inside assignments was rated somewhat more favorably than the women who remained on patrol, whereas the men who were given station assignments were rated less favorably than those remaining on patrol. Women were also assigned to one-officer cars considerably less often than would be expected in their districts, whereas the reverse was true for the men.

Although assignments differed, the types of calls which were handled by men and women were similar, according to the records of the observed tours during the second phase. Of particular concern were the "potentially violent" calls, since there were indications early in the program that some of the (male) partners of the women were behaving in an overprotective manner, and that others were indicating to the dispatcher that they were patrolling alone or with a female, presumably in order to receive additional backup or to
<table>
<thead>
<tr>
<th>TYPE OF POLICE TEAM</th>
<th>NUMBER OF TOURS</th>
<th>TYPE OF INCIDENT AS DESCRIBED BY DISPATCHER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Nothing Found</td>
</tr>
<tr>
<td>One Man</td>
<td>10</td>
<td>0.70</td>
</tr>
<tr>
<td>One Woman</td>
<td>42</td>
<td>1.09</td>
</tr>
<tr>
<td>Two Men</td>
<td>94</td>
<td>0.87</td>
</tr>
<tr>
<td>One Woman/One Man</td>
<td>43</td>
<td>0.56</td>
</tr>
</tbody>
</table>

a Includes all calls for which nothing was found.
b Described on survey instrument as "major disturbances outside of a residence."
c Described on survey instrument as "unspecified with indication of a crime" and as "prowler."
* Significant difference (p < .05)
+ Tends to be significantly different (p < .10)
avoid assignment to dangerous calls. There were observed instances in the second phase of women being taken off dangerous calls (the most blatant instance being an officer-in-trouble call [a male officer's throat had been slit], to which a female officer was initially assigned and then called off, despite her being about three blocks away). However, statistics from the observed tours (1973) indicate that men and women were given similar proportions of dispatched calls for all types of incidents, with the exception of incidents in which there were drunk or disorderly persons, which women were more likely to handle than men (see Table 4). The frequency with which citizens were upset, angry, potentially violent or actually violent upon the arrival of the officers (according to observers), was similar for men and women.

Other expressions of unequal treatment were obtained from interviews of officers, written comments by officers on the attitude surveys, observers' comments, and other unquantifiable sources, and hence, it is not known whether they were frequent or relatively rare events. For example, during the program there were reports of male officers and officials sexually harrassing female officers; subsequent to the evaluation, an article appeared in the Washington Post (1975) charging that "men punish policewomen who don't submit to sexual advances and reward those who do with better assignments and treatment." Occasional instances of men failing to back up female officers were reported (failing to assist them in dangerous situation, particularly when a female citizen was involved, or leaving the scene altogether). On the observed tours, however, additional officers
arrived in incidents handled by male and female officers with equal frequency. There were no observed instances in which officers of either sex failed to receive backup when requested. However, of the incidents described by citizens as having one or more persons present who was "upset, angry, or dangerous" (Service Survey), other police arrived when female officers were assigned to the call only 29% of the time, compared with 42% of the time for male officers. In the majority of these incidents (63%), female officers were patrolling alone, whereas male officers rarely were (8%). Since the sample of incidents is relatively small, however (35 for female and 36 for male officers), it is not known whether this is typical. Results from a sample of observed incidents (1973) handled by officers patrolling alone (19 incidents for males and 93 for females) suggest that female officers patrolling alone request backup in proportion to when it is actually needed, whereas male officers are more likely to request backup even when there is no clear need. According to observers, backup was needed by female officers (alone) 11% of the time, and requested 11% of the time. Male officers patrolling alone needed backup 5% of the time, and requested it 21% of the time. These results are tentative, however, because of the small sample or male officers patrolling alone.

In summary, women had less opportunity to patrol, the sample that remained on patrol was somewhat biased in favor of the men, and women may have received more harassment than men, all of which suggests that the performance findings should be interpreted with caution. These findings are particularly relevant to the administrative
goals, since the pattern is clear -- once the direct orders and program monitoring to ensure equal opportunity are removed, the traditional patterns reappear -- women are placed in stereotypically feminine roles, and different standards emerge for the sexes with respect to assignments. The "better" women are wanted in clerical positions, whereas the less capable men are wanted off the street. Although many women indicated that they enjoyed patrol work and had not requested station duty, societal inequities regarding the responsibility for child care may have influenced these results to some extent. Women on inside assignments (which often have more stable hours) tended to have children more than women on street assignments (65% versus 47% respectively), whereas this was not true for the men (10% on inside and 36% on street assignments had children).

The legal and policy goals were addressed by the following questions:

Did a substantial number of women compare favorably to the average male officer with comparable experience?

How does the average female officer compare to the average male officer in the performance of patrol duties?

Legally, sex may not be a barrier to employment if it can be shown that some members of the sex can perform the job capably, even if the average women is not as capable as the average man. While the courts could, in theory, determine that there was no basis for sex discrimination even if men were found to be significantly "better" than women on all of the performance indicators, it is clearly more
persuasive to show the areas in which women did not differ significantly from the average man. This is true for a variety of reasons: 1) percentage of overlap figures are subject to widely differing interpretations, whereas if it can be shown that men and women performed similarly (no significant differences), then there would be little basis for judging women to be incapable of patrol work. Simply presenting the percentage of overlap carries the risk of bias on the part of evaluators or others interested in the report, without statistical justification. 2) The prevailing technical orientation practically requires statistical analysis for major studies, even if the results could be presented judiciously and clearly in another manner. Proposals without plans for such analysis would generally be rejected, and, even if accepted, carry the risk of the results being dismissed as "unscientific." 3) If women were found to be similar to men in most major respects, this would remove the necessity for developing new selection techniques which would be capable of differentiating between applicants who are likely to be able police officers and those who will perform less than adequately. This is particularly important, since the development of such measures is expensive, requires lengthy validation time, poses administrative difficulties, and, to date, has met with little success in predicting later performance. If important sex differences existed, however, then police departments committed to hiring women might also consider new recruiting, training or supervisory methods, and/or individualizing officer assignments, to emphasize the strong and minimize the weak qualities of officers of both sexes.
In practice, then, the question of sex differences was addressed by first determining whether any differences existed between the average male and female officer, as measured by performance observations, citizen reports of officer performance, ratings by supervisors and departmental performance statistics. When differences occurred on variables which were considered important, and the differences were sizeable and statistically significant, then the extent of overlap as well as the mean differences were reported.

**Performance:** During the course of the performance observations, men generally patrolled with a (male) partner (94 out of 104 tours), whereas women were equally likely to be patrolling alone or with a partner. Patrolmen appeared to be more aggressive in the conduct of their job than patrolwomen. In the first phase of study, when virtually all officers patrolled with (male) partners, observers reported that the women were less likely than the comparison men to "take charge" in situations or to "give instructions." By the second phase, however, the two partners reportedly worked well together, gave instructions to each other fairly equally, took charge with about the same frequency, and were about equal in sharing the driving.

Male officers initiated incidents more frequently while under observation, this difference being particularly apparent when officers who were patrolling alone were compared (male officers patrolling alone initiated 3.20 incidents per tour, compared with 1.05 for females alone, and teams of two males initiated 2.72 incidents per tour, compared with 2.14 for mixed sex teams, both differences being statistically significant \( p < .05 \) ). These incidents consisted
primarily of issuing traffic citations and warnings. Consequently, female officers gave less than half the number of moving traffic citations as their male counterparts. Female officers patrolling alone, however, tended \( p < .10 \) to handle more dispatched calls. Since dispatched incidents take longer to handle, the total time spent handling incidents was similar for men and women.

In both phases, women made less than half the number of felony and misdemeanor arrests and gave less than half the number of moving traffic citations as their male counterparts (see Table 5). There was some overlap in arrest statistics, 20% of the women making as many or more arrests as the typical (median) male officer. As previously noted, women had less opportunity to make arrests, since a disproportionate number were assigned to inside jobs. However, large and statistically significant sex differences in arrests existed even when comparisons were limited to officers who had spent at least 70% of their time (from January 1, 1973 through June 30, 1973), on patrol assignments.

During the second phase, an attempt was made to determine whether men and women made similar types of arrests and whether they were equally credited for arrests. The types of arrests made by officers of both sexes were similar. Due to the small sample of observed arrests, however, the data is inconclusive. Of the 26 arrests made by teams with two male officers regarding credit for arrests, the comparison males were credited with their arrests plus half of the arrests initiated by their partners. Only 7 arrests were ob-
served for male-female teams, and the women were credited for at
least one of the two arrests they initiated (data on one arrest was
missing, so it is not known who was listed as the arresting officer).

<table>
<thead>
<tr>
<th>GROUP</th>
<th>SAMPLE SIZE</th>
<th>PERCENTAGE DISTRIBUTION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No Arrests</td>
</tr>
<tr>
<td>New Women (All Assignments)</td>
<td>70</td>
<td>19</td>
</tr>
<tr>
<td>Comparison Men (All Assignments)</td>
<td>53</td>
<td>4</td>
</tr>
<tr>
<td>New Women Assigned to Patrol</td>
<td>40</td>
<td>10</td>
</tr>
<tr>
<td>Comparison Men Assigned to Patrol</td>
<td>47</td>
<td>0</td>
</tr>
</tbody>
</table>

* Source: Chief's Survey
* Significant @ p < .05

The legality of the arrests was determined by analyzing the
court disposition data from the PROMIS system, which showed that
arrests made by men and women held up equally well in court. A hand
count of the disposition of arrests made by these officers, using
data available in the departmental arrest books, indicated that the
arrests made by women held up better in court than those made by
men (21% of the arrests made by women and 8% of the arrests made by
men resulted in a disposition of "guilty"). This was true even when
arrests having dispositions of "handled by another agency" (e.g.,
military courts, other jurisdictions, etc.), which were more common
dispositions for arrests made by male officers, were all considered
as having a "guilty" disposition. This data was considered less valid than the PROMIS system data, however, and therefore the PROMIS system results were presented as the major findings regarding arrest dispositions (see Tables 6 and 7 for a comparison of the results from these two sources).

Measuring the desirability of arrests was not possible. However, a number of indicators suggested that the women may have had different styles of handling situations which contributed to their low arrest rate. For example, women appeared to believe less in arrests as a tactic for handling situations (as did black males, who made fewer arrests than white males), and there were some indications that they may have provoked arrests less often, or handled provocative situations in a manner which decreased the necessity for arrests. For example, men were significantly more likely to have been charged with serious unbecoming conduct, which is consistent with their generally more aggressive style.

Comments from officials, policewomen and observers may be helpful in understanding some of the possible sex differences in styles of policing which may contribute to the women's low arrest rate (open-ended comments and interview results may be found in Bloch & Anderson, 1974a, pp. 45-52). From one official, "Policewomen tend to be easier to supervise. They tend to be more obedient, they generate fewer citizen complaints. On the negative side, they are harder to motivate to aggressive patrol." Another official stated, "Most women panic easily and have neither the courage nor physical strength to make an arrest unless they are backed up by a man. When
<table>
<thead>
<tr>
<th>MOST SERIOUS CHARGE</th>
<th>SEX OF OFFICER</th>
<th>NUMBER OF ARRESTS</th>
<th>DISPOSITIONS OF CASES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(POLICE OR PROSECUTOR)</td>
<td></td>
<td>Guilty</td>
</tr>
<tr>
<td>Murder</td>
<td>Male</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>Rape</td>
<td>Male</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Assault With a Deadly</td>
<td>Male</td>
<td>53</td>
<td>9</td>
</tr>
<tr>
<td>Weapon</td>
<td>Female</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>Assault on a Police Officer</td>
<td>Male</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Simple Assault</td>
<td>Male</td>
<td>47</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>14</td>
<td>50</td>
</tr>
<tr>
<td>Robbery</td>
<td>Male</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Burglary</td>
<td>Male</td>
<td>11</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>5</td>
<td>40</td>
</tr>
<tr>
<td>Illegal Entry</td>
<td>Male</td>
<td>9</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>Stolen Auto</td>
<td>Male</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>Larceny</td>
<td>Male</td>
<td>31</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>24</td>
<td>38</td>
</tr>
<tr>
<td>Receiving Stolen Property</td>
<td>Male</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Destruction of Property</td>
<td>Male</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Illegal Possession of Weapon</td>
<td>Male</td>
<td>50</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>15</td>
<td>67</td>
</tr>
<tr>
<td>Narcotics</td>
<td>Male</td>
<td>32</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Prostitution</td>
<td>Male</td>
<td>10</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td>Miscellaneous (Gambling, bail violation, forgery, etc.)</td>
<td>Male</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>7</td>
<td>29</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>Male</td>
<td>294</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>121</td>
<td>28</td>
</tr>
</tbody>
</table>

---

* These arrests comprise the total sample of arrests for serious offenses made by all officers under study (86 men and 86 women) from August 1, 1972 through July 31, 1973, as recorded in the Prosecutor's Management Information System (PROMIS). Offenses for which the maximum penalty is less than $300 (primarily disorderly conduct and municipal code violations) and offenses dropped by the officer before reaching the prosecutor are not included in this system.

* Significant @ p < .05
## TABLE 7
POLICE DATA ON ARRESTS AND DISPOSITIONS FOR COMPARISON MEN AND NEW WOMEN

<table>
<thead>
<tr>
<th>OFFENSE</th>
<th>SEX OF OFFICER</th>
<th>NUMBER OF ARRESTS</th>
<th>DISMISSED, NO PAPER OR NOLLE</th>
<th>LOCK-UP</th>
<th>FORFEIT</th>
<th>INDICTED/COURT</th>
<th>OTHER AGENCY</th>
<th>NOT GUILTY</th>
<th>GUILTY</th>
<th>NO DISPOSITION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>Property Crime</td>
<td>Male</td>
<td>24</td>
<td>42</td>
<td>17</td>
<td>0</td>
<td>17</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>28</td>
<td>54</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>29</td>
<td>4</td>
</tr>
<tr>
<td>Disorderly</td>
<td>Male</td>
<td>15</td>
<td>7</td>
<td>0</td>
<td>87</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>21</td>
<td>24</td>
<td>0</td>
<td>62</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Narcotics</td>
<td>Male</td>
<td>12</td>
<td>42</td>
<td>33</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>10</td>
<td>60</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>Gun Law</td>
<td>Male</td>
<td>16</td>
<td>56</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>4</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>Aggravated Assault</td>
<td>Male</td>
<td>8</td>
<td>76</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>12</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>42</td>
<td>0</td>
</tr>
<tr>
<td>Simple Assault</td>
<td>Male</td>
<td>4</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>9</td>
<td>55</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>Fugitive or Warrant</td>
<td>Male</td>
<td>3</td>
<td>33</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>Gambling or Lottery</td>
<td>Male</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>4</td>
<td>75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>Homicide</td>
<td>Male</td>
<td>1</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rape</td>
<td>Male</td>
<td>.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Violation</td>
<td>Male</td>
<td>14</td>
<td>0</td>
<td>7</td>
<td>86</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>Assault on Police Officer</td>
<td>Male</td>
<td>1</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>1</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>All Non-Traffic Arrests</td>
<td>Male</td>
<td>100</td>
<td>34</td>
<td>12</td>
<td>26</td>
<td>13</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>100</td>
<td>40</td>
<td>0</td>
<td>19</td>
<td>13</td>
<td>0</td>
<td>2</td>
<td>22</td>
<td>4</td>
</tr>
</tbody>
</table>

* Significant @ p < .05
working alone, women almost never make any type of arrest. Ninety-nine percent of our women police are just social workers picking up 26 checks a year and alleviating the welfare rolls." (It should be noted that one of the first instances in which a policewoman was commended by the Department involved an off-duty policeman who had shot his wife. While five male officers tended to his wife, the policewoman followed the husband into the next room -- alone -- whereupon he pointed his gun at her chest and she in turn readily disarmed him.) The policewomen themselves provide a different perspective on the issue of aggressive patrol and arrests. Some cited instances of male officers and/or officials inciting the situation after they had calmed it down, while others spoke of using different styles to handle situations. For example, "In one incident, my male partner blew up a minor incident involving a little boy until it became an officer-in-trouble call. Instead of trying to explain to the boy and his parents that the boy was breaking the law, the male officer tried to stop him in a way that antagonized the boy's parents. It developed into a brawl, which infuriated me." Another stated, "I think women are more likely than men to reason rather than to threaten. In one particular incident, two men were fighting, and my male partner goes into the fight with them. The next thing I know, the man was beating up my partner, so I go in it to try to pull the man back. One of the men saw that I was a woman, so he said, 'I don't want to hit a policewoman.'" Another indicated "When I was on the street, I never got overly excited or cursed at citizens, I talked with citi-
zens at their level—and not to them. With my uniform off, I am just like anybody else. I don't need to flaunt my authority the way some men think you have to. Being aggressive is not synonymous with authority."

Observers who were interviewed after completing their last tour related a number of incidents which had impressed them (refer to Bloch & Anderson, 1974a, pp. 18-22, for observer's reports of critical incidents). For example, one observer reported that a policewoman (alone) was handling an incident "beautifully" by persuading an intoxicated man to go home with his wife, until a male officer arrived, put his hands on the drunken man, and started some pushing and shoving -- with the result that the male officer was credited for arresting the man. It should be noted that persuasion was not the only approach used by policewomen. For example, a male observer (a police officer who had initially been somewhat leary of women's ability to handle violent situations) indicated that he had changed his mind after seeing a policewoman (alone) flip a 250-pound, belligerent, intoxicated man over her shoulder and handcuff him.

Departmental priorities about arrests were unclear. Although the department does not condone quotas for arrests or traffic violations, some supervisors are known to have informal quotas, which, if not met, may result in undesirable beats or assignments. On the other hand, there were known instances of male officers who were considered to be overly aggressive in making arrests being temporarily assigned to the station to "cool off." The correlational analysis of background and performance variables revealed that, for female
officers, there was no relationship between the number of misdemeanor arrests and any background or performance variables. For both sexes, departmental ratings appeared to be based on factors other than arrest statistics. These findings suggest that arrests may not be a high priority for the department as a whole. It is not known whether women could have made as many arrests as the men had they been required to do so.

Given the overlap in arrest statistics, the difficulties in interpreting arrest statistics, the problems posed by encouraging aggressive patrol, and the fact that the department places no explicit priorities on making arrests, it would be difficult to conclude that women should be excluded from patrol because of these differences alone.

Specific officer actions and citizen reactions were reported in detail by observers, and to some extent by citizens involved in these incidents. Observers reported that male and female officers were similar in their attitudes and level of respect towards citizens, and that citizens who were directly involved in these incidents, as well as spectators were similar in their reactions to officers of both sexes. When the citizens involved in these incidents were interviewed later, they rated the officers even more favorably on these dimensions, with no sex differences. Police teams with two male officers were more likely to take "action" (do an investigative search, threaten the use of force, or arrest), than mixed sex teams. Other officer behaviors (conversation, give advice/take a report, or actually use force), did not differ significantly by sex, and citizen
reactions to all of these specific officer behaviors, including "action," did not vary by sex of officer. In incidents in which citizens were upset, angry, potentially violent, or actually violent, both male and female officers tended to reduce the level of tension, or to leave the scene with fewer citizens upset than upon entering. Citizens who were interviewed reported a high level of satisfaction with the way the officer(s) had handled the incident, regardless of sex of officer(s). They indicated that if they needed police service in the future, they would slightly prefer the same officer(s), and, if two different officers arrived, that they would have no preference for all male or mixed sex teams. In addition, they supported the idea of women having the same opportunity as men to be police officers, and those who had received police service from a policewoman indicated that their experience had given them a slightly more favorable attitude towards policewomen. Citizens were also more likely to indicate that someone (a citizen) present in the incident had been angry than were the observers. According to citizens, someone was angry in 43% of the incidents female officers handled, compared with 31% of the incidents male officers were assigned to. In all incidents handled by female officers, the citizen was angry before police arrival, whereas in 26% of the incidents male officers handled, this occurred after their arrival (significant @ p < .05). Female officers handled a significantly larger percentage of incidents in which citizens reported that someone present was "upset, but not angry" (62% compared with 40% for the males), this difference being even more marked when females patrolling alone were compared with males.
patrolling alone (62% vs. 20%). In over 90% of the incidents in which a citizen was upset, the individual was upset prior to police arrival, regardless of the sex of the officers handling the incident.

Officer ratings by supervisors differed on the three rating forms (the Departmental Rating Form, the Official's Survey and the Chief's Survey rating forms). There were no sex differences on any of the categories listed on the Departmental Rating Form (see Table 8). Women average 5.3 and men 5.5 in overall rating, on a scale where ratings of 4 to 6 meant "effective and competent," six being the high end of the scale. When overall ratings of men and women who were still on street assignments by the end of the second phase are compared, however, women average 5.1 and men 5.6, a slight but statistically significant difference. This difference is accounted for by the transfer to inside assignments of 25 women whose overall rating was above average for women (5.5), and of 9 men whose overall rating was below average for men (4.9).

On the Chief's and Official's survey ratings, the men fared better than the women, primarily in situations involving violence. Although the women were rated as average or above average on all categories of the Chief's Survey, men were rated higher than women on "general performance of street patrol," "protecting a partner from violence," "handling a public fight," and "handling a disorderly male." There were no statistically significant sex differences on "dealing with the public," "handling an auto accident involving an injury," "making a crime report," or "handling a disorderly female." On the (anonymous) Official's Survey, women were rated as less com-
### TABLE 8
DEPARTMENTAL PERFORMANCE RATINGS\(^a\)

<table>
<thead>
<tr>
<th>TYPE OF RATING</th>
<th>AVERAGE RATING(^b)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NEW WOMEN</td>
<td>COMPARISON MEN</td>
<td></td>
</tr>
<tr>
<td>Bearing and Behavior</td>
<td>5.4</td>
<td>5.7</td>
<td></td>
</tr>
<tr>
<td>Human Relations</td>
<td>5.5</td>
<td>5.7</td>
<td></td>
</tr>
<tr>
<td>Learning Ability</td>
<td>5.4</td>
<td>5.5</td>
<td></td>
</tr>
<tr>
<td>Knowledge and Skill</td>
<td>4.9</td>
<td>5.2</td>
<td></td>
</tr>
<tr>
<td>Acceptance of Responsibility</td>
<td>5.1</td>
<td>5.4</td>
<td></td>
</tr>
<tr>
<td>Written Expression</td>
<td>5.4</td>
<td>5.4</td>
<td></td>
</tr>
<tr>
<td>Oral Expression</td>
<td>5.4</td>
<td>5.3</td>
<td></td>
</tr>
<tr>
<td>Performance of Duty</td>
<td>5.1</td>
<td>5.5</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>5.3</td>
<td>5.5</td>
<td></td>
</tr>
</tbody>
</table>

\[^a\] These ratings are given to all officers at the end of the probationary year; the ratings are taken into consideration for step increases.

\[^b\] 1 = unsatisfactory, 2 = marginal, 3 = below average, 4-6 = effective and competent, 7-8 = excellent, 9 = exceptional, 10 = outstanding.
petent than men on their ability to handle "domestic fights," "street violence" and "general competence," and equal to men on ability to handle "upset or injured people." When captain's and lieutenant's ratings (1973) are considered separately from sergeant's ratings, women differed from men only on ability to handle "domestic fights." Captains and lieutenants improved their ratings of women in all categories from 1972 to 1973, whereas sergeant's ratings were unchanged.

Despite the differences between the rating forms, all three forms were positively correlated. The overall rating on the Departmental Rating Form was one of three major performance variables which was correlated with a number of background and (other) performance variables, in an attempt to provide a picture of the male and female officer who was considered favorably by supervisors. Both male and female officers with better departmental ratings were found to have higher average training academy scores, fewer charges of unbecoming conduct and more written commendations from citizens. White officers, regardless of sex, had better overall ratings than black officers. Women with higher departmental ratings tended to be more educated, to have higher Civil Service scores, to be less likely to have resigned, to be more likely to have spent a large percentage of their time on station duty, to have a less friendly attitude towards dangerous or angry citizens, and to elicit a more favorable reaction from citizens who had been victims of a crime than women with lower departmental ratings. Men with higher departmental ratings tended to
be older, married, were less likely to have been arrested for a crime prior to entering the department, were involved in fewer auto accidents, and received more favorable comments from citizens than men with lower overall ratings.

The remaining performance variables relate primarily to administrative concerns, rather than to performance which may seriously affect the question of whether women can perform on patrol in a manner satisfactory to the department and the community. These include sick leave, injuries, mild misconduct, resignations, and driving skills. Contrary to the opinions of male officers and officials, men and women did not differ in the number of days of sick leave taken or in the frequency with which they used sick leave. Women were, however, somewhat more likely to be assigned to light duty as a result of injury. Nevertheless, the average number of injuries sustained was similar for both sexes, even when women who were primarily assigned to patrol duties were compared with men who were primarily on patrol assignments, and these injuries did not cause women to be absent from work more often than men. It seems likely that the light duty statistics reflected sex differences in the treatment of injuries, rather than in incidence. Some women may more readily request light duty following an injury than would men with similar injuries, and some supervisors may more willingly grant light duty to women with relatively minor injuries than they would to male officers. Men and women had an equal number of driving accidents during the course of duty. Men were more likely to be engaged in mild (as well as serious) misconduct, mild misconduct including
being overweight, improper appearance or equipment, improper record-keeping procedures, and non-attendance or sleeping in required classes. Women were more likely to have been noted for being tardy, with no indication that they had been involved in any more serious offense. By October, 1973, 12 women (14%) and 11 men (13%) had resigned, 4 of these men and 2 of these women having been terminated "in the best interests of the department."

The attitudes of the community and of police personnel toward policewomen serve the persuasive and administrative goals, and may address the legal goals of such attitudes impinge upon the women's opportunity to perform. The following question address these goals:

What are the attitudes and beliefs of police officers and officials about female patrol officers?

How does the community feel about women on patrol?

Favorable or unfavorable attitudes toward policewomen may influence departmental policy, regardless of the legal situation. For example, prior to the D.C. evaluation, a police department which had planned to hire policewomen decided against this after the policemen's wives association strongly opposed the plan. Another department indicated that they would not hire policewomen because they felt the community would not accept it. The attitudes of coworkers and supervisors, as well as community attitudes, are part of one's work milieu. If there are major differences between the way men and women are treated by those they must work with, then this may affect their performance. In addition, attitude changes over time provide one type of indicator of whether the women are gaining or losing esteem.
as a result of their performance. Finally, the attitudes of the officers under study toward their job may help explain differences in job performance.

**Police attitudes:** Attitudes of male officers and officials were quite negative towards policewomen before and during the study period. Officials were slightly more positive than officers during the second phase, but by and large it was clear that women were expected to perform more poorly than men, particularly in situations involving violence. For example, when officers and officials were asked to indicate the percentage of male and female officers who they believed would be satisfactory in handling 13 different patrol situations, male officers indicated that more men than women would be capable of handling 10 of the 13 situations (see Table 9). The only patrol skill which they felt women had an edge over men in handling was "questionning a rape victim." Policewomen, however, believed that their patrol skills were, for the most part, as good as the men's, and in a few cases, better. Officials were neither as sceptical as the male officers nor as positive as the female officers. When asked their opinion about the effect of a male-female team versus an all-male team of officers on two potentially violent situations (three people fighting and a domestic fight in which the man is holding a gun at his side), officials and male officers felt the presence of a policewoman would be slightly detrimental (i.e., the situation would escalate), whereas female officers felt a male-female team would be slightly advantageous. Male officers and officials also felt that both the female officer and her (male) partner
### TABLE 9
PERCENT OF OFFICERS WITH ONE YEAR'S EXPERIENCE
BELIEVED "SATISFACTORY" AT PATROL SKILLS

<table>
<thead>
<tr>
<th>SKILLS</th>
<th>SEX OF OFFICERS EVALUATED</th>
<th>GROUP GIVING RATING</th>
<th>SUMMARY OF VIEWS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Officials (N=84)</td>
<td>Patrolmen (N=82)</td>
<td>Patrolwomen (N=14)</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>Questioning a Rape Victim</td>
<td>Women</td>
<td>66</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>39</td>
<td>68</td>
</tr>
<tr>
<td>Writing Reports</td>
<td>Women</td>
<td>67</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>60</td>
<td>68</td>
</tr>
<tr>
<td>Arresting Prostitutes</td>
<td>Women</td>
<td>48</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>45</td>
<td>58</td>
</tr>
<tr>
<td>Settling Family Disputes</td>
<td>Women</td>
<td>57</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>59</td>
<td>68</td>
</tr>
<tr>
<td>Getting Information at Crime Scenes</td>
<td>Women</td>
<td>50</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>54</td>
<td>68</td>
</tr>
<tr>
<td>Handling Disorderly Females</td>
<td>Women</td>
<td>49</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>57</td>
<td>63</td>
</tr>
<tr>
<td>Cruising Around and Observing</td>
<td>Women</td>
<td>55</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>65</td>
<td>73</td>
</tr>
<tr>
<td>Handling Traffic Accidents</td>
<td>Women</td>
<td>61</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>71</td>
<td>78</td>
</tr>
<tr>
<td>Handling Victims of an Armed Robbery</td>
<td>Women</td>
<td>55</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>66</td>
<td>67</td>
</tr>
<tr>
<td>Dispersing a Group of Noisy Juveniles (ages 9-12)</td>
<td>Women</td>
<td>49</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>67</td>
<td>73</td>
</tr>
<tr>
<td>Handling Threatening Situations Where Someone Has a Knife or a Gun</td>
<td>Women</td>
<td>32</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>60</td>
<td>68</td>
</tr>
<tr>
<td>Handling a Down and Out Drunk</td>
<td>Women</td>
<td>48</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>81</td>
<td>81</td>
</tr>
<tr>
<td>Handling Disorderly Males</td>
<td>Women</td>
<td>35</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>69</td>
<td>75</td>
</tr>
</tbody>
</table>

* The percent of officers rated "satisfactory" was indicated on the questionnaire by placing a mark on a line labelled zero percent on the left end of the line, 50 percent in the middle and 100 percent on the right end. Answers were coded into 9 categories (0-9), and average scores were converted into equivalent percents.

* Significant @ p < .05
would be somewhat more likely to be injured in both situations than if two male officers had responded, whereas women felt the presence of a woman would slightly improve the safety of both officers. Police officials and male officers indicated a definite preference, and female officers a slight preference, for a male partner. As an indication of the strength of sex bias in the department, respondents were asked the following question: "If you told your fellow patrol officers that you had a female partner who you thought was a better police officer than many male officers, what percent of them do you think would believe you?" Male and female officers indicated that about 40% would believe them, and officials felt about half would believe. When asked whether women should be a regular part of the patrol force, 48% of the male officers, 55% of the officials and 79% of the women said "yes," although some felt that the duties or numbers of women should be restricted. Overall, male officers changed their attitudes toward women very little, and officials were only slightly more favorable from 1972 to 1973.

Policewomen surveyed reported receiving more cooperation from the public in obtaining information about a crime than did male officers, and also reported fewer insults, fewer attempted injuries, and more compliments from the public within the previous month's time. (Differences were not statistically significant for any but the community cooperation variable during the second phase, although all were during the first phase, due in large part to the small sample of female respondents in the second phase.)
In order to assess sex role stereotypes, respondents were asked to indicate the percentage of male and female officers who they felt possessed 12 traits which were selected because of their similarity to characteristics commonly attributed to men or women (Broverman et al., 1970), as well as for their relevance to patrol work. Respondents were also to indicate up to six of these traits which they considered very important to a good patrol officer. Officials, males and females agreed that being "calm and cool in tough situations" was the most important trait, followed by "thinks and acts decisively," and "observant." Male officers thought that men and women were equally likely to be "calm and cool in tough situations," officials felt that men had an edge, and women felt they held a slight advantage. Male officers credited women with being slightly more "intelligent" and "understanding," but officials and male officers agreed that men were more likely to "think and act decisively," to be "observant" and to be "emotionally stable." Female officers generally felt patrolwomen were as likely or more likely than patrolmen to possess all of these traits.

Nearly all female officers reported that they talked regularly with a sergeant (93%), compared with 62% of the men. However, men and women differed in their perceptions of official's actions. Women felt that officials were more critical of policewomen than policemen and that they gave the men more useful job suggestions than the women, whereas male officers felt there was no sex differences on these variables. Policewomen felt that women who performed as well
as men received worse job ratings, while men felt that they were rated somewhat better. Given the conflicting opinions, it is difficult to ascertain whether women were treated and rated fairly by officials.

The attitudes expressed during the second phase of study were by and large very similar to those in the first phase and to the results from a smaller sample of male officers pretested before the women went on patrol. Hence, these attitudes do not necessarily reflect the officers’ experience with policewomen.

Community attitudes: A representative sample of citizens who were residents of the E and C districts indicated their general approval of: women on patrol, equal opportunity for equally qualified women on the force, and the women's rights movement. These attitudes were shared by all sub-groups, including men and women, black and white respondents, and residents of the E and C districts, and business persons. When questioned about specific patrol situations, citizens generally believed there was little difference between the performance of policemen and women in most situations. In two "violent" situations, however (a street corner brawl and a riot), the average citizen felt that an all-male team would have an advantage over a mixed-sex team. Respondents gave women an edge over men in investigating a crime scene, and also felt a male-female team would be somewhat more effective in handling a fight between a man and a woman. Citizens, and particularly black citizens, felt policewomen would show more respect than policemen for people in their neighbor-
hood, and in turn felt that policewomen would be treated with slightly more respect from the community than male officers.

Most citizens interviewed in 1973 (70%) had seen or heard about policewomen through the media during the previous six months. About one-third of these citizens felt they had a "more favorable" or "much more favorable" attitude toward policewomen as a result, virtually all of the remaining citizens indicating that their attitudes had not changed. One-third of the citizens interviewed in 1973 had seen a policewoman handling an incident, had talked with a policewoman, or knew a policewoman in the six months prior to being interviewed (about twice as many males as female indicated this). As a result of this contact, 44% reported a "more favorable" or "much more favorable" attitude towards policewomen, 51% indicated no change in attitude, and 4% had a more negative impression of policewomen. The level of personal contact in 1973 was higher than for any of the three previous administrations of this survey, and the positive effect of this contact was also higher than it had been in earlier surveys.

The research goals were addressed by all of the findings, since the results contributed to the state of knowledge in the field. Questions specific to the research goals concerned the following:

**How methodologically sound was the evaluation?**

The study design, the experimental procedures, the reliability and validity of the measurement instruments, the analytic techniques utilized, and the accuracy in drawing inferences from the data are
all important components of this question. If the study is seriously
deficient methodologically, then the results cannot be relied upon
for the other goals of the evaluation.

Methodological soundness: The relative importance of the find­
ings is only in part determined by the significance of the results
and the soundness of the experimental procedures. The relevance of
the data to the goals and values of the users of such information
weigh heavily in such a determination. Thus, knowledge of police
department functioning and priorities and the legal requirements for
evidence in sex discrimination cases are as important in determining
which measures should be regarded as the more accurate and relevant
indicators as are the statistics and knowledge of the problems en­
countered in administering and interpreting the measures.

The limitations of each of the types of measures, the specific
problems in instrument design and administration, and the rationale
for considering the reliability and validity of these measures in a
different light than would be appropriate for laboratory research have
been discussed in detail in the methodology section. The fact that
these performance measures were carefully constructed in a standard
format, that the items to be measured were generally straightforward
and subject to little differential interpretation, that the items
were not attempting to measure constructs beyond the questions them­selves, that similar performance items were often gathered in a
variety of ways, from a variety of perspectives, that most of these
measures were administered at least twice, with similar results, and
that a review by experts in the field found these measures to be ade­
quate supports the author's assertion that indices of reliability and validity were not particularly important, given the goals of this evaluation. Performance measures were considered by the researchers to be by far the most important measures in this study, and the performance measures which were the most direct and representative measures of the job were considered the most valid, i.e., the performance observations. Citizen satisfaction with service and departmental ratings were considered to be the most valid "indirect" measures of performance.
DISCUSSION

The most important question for the policewoman evaluation was whether or not women could handle the job of patrol as well as men with similar qualifications and experience. This question was particularly critical, since a major goal of the evaluation was to provide information which would be useful for the legal determination of an essentially moral issue. That is, the evaluation was designed to provide empirical evidence about the relative capabilities of women on patrol, in order to bear upon the issue of sex discrimination in police departments. The EEOC Guidelines on Discrimination Because of Sex (1972) clearly limit the grounds for exemption (from the requirement to hire women on an equal basis with men) to cases in which there is an overriding business necessity\(^1\), and the employer has made

\(^1\)The EEOC is issued a document entitled "Questions and Answers Concerning the EEOC Guidelines on Discrimination Because of Sex." In response to the question of the conditions under which an employer may establish a "business necessity" defense for sex discrimination, they quote the court decision in Robinson vs. Lorillard Corp., 444 F. 2d 791, 798 (4th Cir., 1971), a case charging race discrimination. "The test is whether there exists an overriding legitimate purpose such that the practice is necessary to the safe and efficient operation of the business. Thus, the business purpose must be sufficiently compelling to override any racial impact; the challenged practice must be sufficiently compelling to override any racial impact; the challenged practice must effectively carry out the business purpose it is alleged to serve; and there must be available no acceptable alternative policies or practices which would better accomplish the business purpose advanced, or accomplish it equally well with a lesser differential racial impact."
reasonable attempts to develop selection criteria which would permit the hiring of qualified members of both sexes, or when a particular sex is needed for the purpose of authenticity or genuineness (e.g., acting or modelling jobs). With respect to policewomen, this legislation implies that it must be shown that women are not capable of performing patrol functions, and that there is no reasonable way in which women could be hired or trained in these duties without seriously undermining the effectiveness of the police department.

This legislation is derived from a moral, rather than empirical origin (i.e., the position that "all men [and women], are created equal"). The historical fact that women have been denied full access to employment opportunities in jobs which are stereotypically masculine, many of which have high status and/or which are highly remunerative in comparison with the opportunities typically available to women, has probably influenced legislators in their decision to enact and enforce legislation prohibiting sex discrimination, and has certainly contributed to the pressure for legislative action from women's advocacy groups. (For example, Bem and Bem [1972] report that fewer than 1% of all working women work in occupation which most Americans consider "professional." Even gifted women are not immune—Terman and Oden [1959] found that 86% of the men studied in the "genius" range achieved prominence in professional and managerial positions, compared with 61% of the women (at age 44) being full-time housewives. Of the women who worked, only 11% were in comparable positions. The salary differential is similarly dramatic—the average full-time working woman can
expect to earn slightly more than half (58%) the salary of the average male worker [Bem & Bem, 1972]. The two major reasons cited for this are that, in every category of occupation, women are employed in the lesser-skilled lower-paid positions, and that women receive unequal pay for equal work.) Nevertheless, the legal grounds for considering departmental practices to be discriminatory must be based upon demonstrated performance, rather than historical inequities. This evaluation thus had major implications for the administration of social policy, and in this sense differs from most evaluation research, which tends to have narrower goals.

The policewomen evaluation also differed from the typical program evaluation in a number of other respects, which clearly influenced the focus, standards and conduct of the study itself. 1) The social policy issue addressed changes in culturally defined sex roles in an occupation which has a long tradition of being seen as "men's work" by both police personnel and the larger culture (see Astin, Suniewick & Dweck, 1974 for a bibliography on women and careers). The legislative changes regarding sex discrimination and the changing attitudes of some segments of the community towards women's rights were in distinct violation of the cultural practices of sex-role distinctions and sex discrimination. Thus, it was more evident in this program than in most that implementation was going to be opposed by the strength of tradition and prejudice, and that much effort would be required.

2) To a degree, the results of the study were intended to be used as evidence in judicial proceedings. In such proceedings, the political, economic and social costs and benefits of program implementation,
which ordinarily are quite important in determining the ultimate impact of a program, were largely irrelevant to the court decisions to be made from this evaluation. The success of the program, as measured by various performance criteria, which normally plays a small role in determining the evaluation impact, was paramount to this effort (refer to Appendix B for a more thorough discussion of the objectives of program evaluations).

Police administrators in departments across the nation were the other major audience. While their role is critical in the successful implementation of such a program, it is inextricably tied to the position taken by the courts, since voluntary implementation is considerably more likely to occur in such a controversial program when there has been legal precedence for compliance. As more departments are required by the courts to hire women for patrol duty, the practice will become less novel, and hence the culturally defined sex roles will become less distinct, and the reactions to the violation of norms less acute. The experiences of other departments will tend to reduce unwarranted fears, and lessen the outrage about breaking with tradition in their own departments. Finally, legal precedence will increase the threat of legal action, which may in turn increase the likelihood of departments dealing with sex discrimination to avert such action.

3) The decision-makers within the judicial system were not invested in the program per se, but neither were they neutral in their stance towards the evaluation results. In view of the strong legislative and judicial commitment to the abolition of sex discrimination, it was not necessary to show that women were equal in all respects to men in
the performance of patrol duties, but rather to indicate whether or not there was compelling evidence that most women could not handle patrol work. Since there was a strong commitment to the amelioration of sex discrimination, less stringent evidence of a desired effect was necessary than is required by a purely academic study.

Even if more substantial sex differences had been obtained, this would not necessarily reflect inherent differences in the capabilities of men and women. It was recognized that the cultural conditions under which the women in the study were raised probably influenced the study results to some (unknown) extent. For example, women are generally raised to be considerably less physically aggressive than men and to take initiative less frequently than men (Maccoby, 1966); the conflict engendered by adopting a "sex-inappropriate" role (i.e., police work) must impose an additional burden upon the women not experienced by men. As one official stated, "in our society, men are trained from infancy to take the dominant role in physical situations. Women are trained to let them and only step in when they have to. When a woman goes into patrol work not only does she have to accomplish a cultural role-reversal within herself, but she also has to demonstrate her ability to do so to and for her comrades and the citizens when she meets the appropriate situations" (Bloch & Anderson, 1974a, p.46). The expectations held by both men and women about women's abilities on "competence" generally fall far short of the expectations for their male cohorts (Maccoby & Jacklin, 1974; Goldberg, 1968). Sex role conditioning and fear of success appears to be part of this configuration (Horner, 1969; 1970). As cultural conditions change, including both the beliefs about
the abilities of women as well as the actual roles which women perform, so may the evaluation results.

Cultural conditions, such as the attitudes of the police department and the community towards policewomen, and the obstacles to program implementation, are in and of themselves irrelevant to the legal issues (EEOC Guidelines, 1972), except insofar as it can be shown that they biased the findings. In most respects, the prevailing conditions tended to bias against the women's success. The attitudes of male officers and officials in the department were generally hostile to the policewomen before and during the experiment. Instances of harrassment or failure to support the policewomen were noted, as were some overly protective actions. Women complained of unequal treatment by supervisors and coworkers. Expectations for policewomen among the rank and file were generally quite low, despite the administration's support and commitment to the program. Women were not given an equal opportunity to patrol, and were not assigned permanent scout cars. The community expected women to be somewhat less able in handling violent situations, but otherwise was favorable in their attitudes. Given the total conditions under which the women worked, with the known influence of expectations upon behavior (Rosenthal, 1966), the fact that relatively few differences in performance were found is surprising, and should be viewed as an additional indication of women's ability to adequately perform the job of patrol officer.

It should be noted that the employment of more men than women is not necessarily discriminatory, provided that departments were able to develop selection techniques which were generally considered to be
relevant to officer performance, i.e., which selected out the applicants having a high probability of performing on a substandard level. If these selection techniques resulted in a disproportionate number of male officers being selected, this might not be construed as discriminatory if the test was not designed, intended or used to discriminate on the basis of sex. Although there have been some ambitious attempts to develop selection techniques which would not discriminate on the basis of race (Baehr et al., 1971), the difficulties inherent to the development of measures which are practical to administer and which predict to actual performance on the job have yielded disappointing results to date (Kent & Eisenberg, 1972). Even if such tests were to be developed, the variations between police departments in standards of acceptable performance, or even within the same department over time, would probably preclude such tests being generally applicable.

The evaluation results generally answer the major question in the affirmative. As summarized in the final report, "The men and women . . . performed patrol work in a generally similar manner. They responded to similar types of calls for police service while on patrol and encountered similar proportions of citizens who were dangerous, angry, upset . . . or violent. Both men and women officers were observed to obtain similar results in handling angry or violent citizens. There were no reported incidents which cast serious doubt on the ability of women to perform patrol work satisfactorily, and (there were) . . . reports of some incidents in which individual women performed quite well in difficult circumstances . . . ratings, reflecting overall departmental judgment of the performance of an officer, indicate equal
overall satisfaction with officers of both sexes. Citizens, either interviewed for the purposes of this study or observed at incidents involving police officers, showed similar levels of respect and generally favorable attitudes toward officers of both sexes. No differences were discerned between male and female officers in their levels of respect and generally favorable attitudes towards citizens. Male officers were found more likely to engage in serious unbecoming conduct" (Bloch & Anderson, 1974a, p. 2).

There are some qualifications to these findings, however. As a group, women made fewer arrests and gave fewer traffic citations (about 20% of the women made as many or more arrests as the typical (median) male officer). The fact that the women had less opportunity than men to make arrests because of a disproportionate number being assigned to inside jobs tends to soften, although by no means totally explain, the differences in arrests and traffic citations. Also, women were somewhat more likely to be assigned to light duty as the result of injury on the job. This finding, however, appeared to be largely due to differential treatment of injuries sustained by men and women, rather than because women were more injury-prone. Finally, women were rated less favorably than men on anonymous rating forms developed for the evaluation, in contrast to the Departmental Rating Form results.

The utilization of the results of social science research for social change efforts has precedence in the Supreme Court decision regarding racial discrimination, in which a moral stance was supported by empirical evidence, resulting in large scale enforcement of that moral position. The problem encountered in attempting to base legal
decisions on research data alone is that in seeking to support a moral position, it often fails to address the social, political, economic and psychological costs of implementation. The danger in this is that the results may be overvalued, and, when difficulties arise, the reactions to the failed expectations may seriously damage the intent of the program or policy. With respect to the policewomen evaluation, it should also be noted that departments vary in their styles of policing—police performance which is acceptable for one style of patrol may be unacceptable for another. Therefore, it is important to evaluate the program problems and to anticipate the likely results, both positive and negative, of large-scale implementation.

The advantages of hiring women on a equal basis with men for patrol work might include the following:

1) Enlargement of the supply of personnel resources, which would permit hiring greater numbers of police personnel from the surrounding community, thus decreasing the cost of recruiting, reducing police turnover, and resulting in a police force which would be familiar with the problems of the community and more representative of the racial and sexual composition of the city. Because of the larger sample of applicants, police departments could be more selective in their choice of officers, thus upgrading the general quality of the force.

2) A department with a substantial number of policewomen may be less aggressive than one with only men. Women act less aggressively and they endorse aggressive police tactics less than male officers. The presence of women may stimulate increased attention to ways of avoiding violence and tempering violent situations without the use of force.
3) If the District of Columbia is representative of other communities, hiring women for patrol may improve relations between the police department and the community. Citizens generally supported the concept of equal employment opportunities for women, both in the police department and in general, and had few reservations about women's abilities. In some respects (primarily in the area of police-citizen attitudes of respect), they felt female officers would be slightly preferable to male officers. Women were also found to be less likely to become involved in serious unbecoming conduct, which can damage community relations.

4) The presence of policewomen may be advantageous in a number of situations, such as family disputes, in which another woman is involved. Further research on this issue is needed, however, before this can be definitively stated.

5) Policewomen may have a positive impact upon the department. In the words of one official, "I find that many male officers are of substandard quality in the areas of intelligence, fundamental knowledge (such as the ability to write, spell, punctuate, etc.), commitment to community, work habits and dress. If the female officers actually worked at being sloppy, illiterate and inefficient they could not be worse. To be brief, I feel the intelligence and commitment of several female officers to be in refreshing contrast to the crybaby attitude of many males. I believe in some instances the female officers shame their male counterparts into doing some work, cleaning up their appearance, improving their vocabulary, and increasing their knowledge of fundamentals such as report writing, interviewing, etc. If the females
served no other function—and they do—this would justify having the female officer in the patrol division" (Bloch & Anderson, 1974a, p. 45).

6) Finally, hiring women brings the department into compliance with federal anti-discrimination legislation, which may avoid the possibility of costly and disruptive lawsuits being levied against the department.

The disadvantages of hiring policewomen include the following:

1) Introducing such a program may be disruptive to the functioning of police departments. There will doubtless be opposition from a sizeable proportion of officers and officials within the department to this change; as a result, low morale, increased problems in department management, and possibly poorer service to the public may occur during the period of adjustment. Men may feel less secure patrolling with women than if they were working with another male officer. In departments in which the police management is unfavorably disposed to women, this opposition may be intensified. The program may falter if management does not develop monitoring and enforcement techniques to ensure that women are treated similarly to the men, with the result being that women will likely be placed on assignments which fall within traditional sex role definitions within the department. If opposition is intense, monitoring of daily assignments, dispatch procedures, and interactions between female officers and their partners may be necessary to counteract harrassment and/or overly protective actions towards men. However, such monitoring requires extra time spent by personnel within the department for this purpose, which is somewhat disruptive of normal activities. Monitoring may engender resistance both because of the effort required and because it represents a violation of privacy—
departmental observation of officers' activities is generally unpopular, and it is difficult to allay fears about the possible uses of the information gained from such efforts. If outside evaluation is undertaken, this may increase both the disruptive effects and the fears of observation.

2) There may also be some problems with the women, unless the program is administered properly. Resistance may be encountered from the women if the department's expectations regarding duties and performance standards are not clearly communicated. This was particularly evident in the D.C. program with the women who were reassigned to uniformed patrol duty after having served in more traditional roles up to that time. Although they seemed to weather the change rather well, there was a great deal of conflict and hostility from the women in the initial stages, primarily because their job expectations had never included patrol work, and even in their retraining sessions, for patrol work, some police officials indicated that it was unlikely they would be assigned to patrol duty. In addition, police departments are para-military organizations, which are less familiar to women in this culture, and hence they may have more difficulty adapting to this style of operation.

3) If the program is implemented when the issue of policewomen is still volatile, there is a risk of unfavorable community reaction, particularly if unfavorable and dramatic incidents occur early in the program (such as incidents in which a policewoman is killed, or performs very poorly in an incident which is highly visible, sensitive, and publicized); policemen's wives associations or other police associations strongly
oppose the program in an organized fashion; or if the media gives unfavorable coverage (even in the absence of dramatic incidents), which adversely affects community attitudes. This risk is probably fairly minimal, however, since the communities surveyed to date have been favorably inclined towards policewomen, and, even after the death of the first policewoman, followed by an incident in which a policewoman killed a suspect for very questionable reasons (both occurring in D.C. subsequent to the evaluation), the reaction of the media was surprisingly "positive"—these events being construed as evidence of women's willingness to handle dangerous situations alongside the men (Armat, 1975; Morgan, 1974).

It should be noted that the D.C. evaluation took place under relatively favorable circumstances. The police administration was extremely supportive, the initial monitoring efforts were thorough, attempts to subvert the program were quickly and effectively dealt with (during the experimental period), the media was generally favorable or neutral in their coverage, and no critical incidents occurred during the course of the evaluation which seriously cast doubt upon the program. Other departments may not have such favorable circumstances, in which case it will be more difficult for the women to prove themselves. Few police chiefs are as firmly supportive as Chief Wilson in this regard, which was an important factor in the women even being permitted the opportunity to demonstrate their abilities. Other departments, however, will not have to suffer the burden of being a test case, during a period in which there was virtually no precedent for such a program. As such, there was very little support from other
police departments, the community and police personnel were psychologically unprepared for this sudden change in the role definitions of a police officer, and the publicity surrounding the program was intense. As hiring women for patrol duty becomes more routine, the pressures felt by the D.C. Department should lessen.

The values of the particular department should be considered in determining the extent to which the implementation of the program will conflict with their usual practices. James Q. Wilson (1968) identifies three major styles of policing, which may be described as: 1) the watchman style, which emphasizes maintaining order rather than law enforcement, 2) the legalistic style, which emphasizes arrests and law enforcement, and 3) the service style, which emphasizes pleasing the public and informal sanctions such as warnings, rather than formal sanctions such as arrests. A department which values arrests and a generally aggressive style of policing will have to weigh the benefits against the possible costs of poor community relations, since aggressive patrol behavior is not always confined to situations where such action is warranted. Skill in interpersonal relations (American Bar Association, 1973), and a better working knowledge of the law (Bozza, 1973), may be more important in solving problems of a potentially violent nature. If a department feels that arrests and moving traffic violations are important performance criteria, this can be explicitly stated and/or required, although this may encourage misplaced aggressiveness. Given that there is no formal requirement for arrests in the D.C. Department, it cannot be assumed that women would be unable to make as many arrests as the men had they been encouraged or required to do so.
IMPLICATIONS

The factors which contributed to the utilization of this study, without which it might have shared the obscurity of many other evaluations, are numerous. Some of the more salient factors will be presented, including the study's social relevance or "timeliness," the study design and methodological strategies, the characteristics and roles of the study participants, and the advance planning for impact, including the role of the media.

Social relevance

One of the most important factors in enhancing the impact of the study was that it dealt with an issue which was politically and socially timely. Feminism was a widely debated topic, and sex discrimination was finally being considered seriously by the courts and federal funding agencies, to the point where pressure, and occasionally economic sanctions, were beginning to be applied to institutions receiving large amounts of federal funding, such as universities, when they continued to be remiss in their hiring and promotional practices. The political climate was such that awareness of sex discrimination had cut across all levels of society, and administrators and employees were becoming increasingly uncomfortable with practices which would have been given little thought even five years before. Although many professions and other occupations had voluntarily lowered their barriers against
women, the archetypically masculine fields, such as the military and police and fire agencies, had barely responded, seeming to be just about as staunchly male oriented as ever in their policies and attitudes. Women, however, were ready to respond to widening opportunities in these areas, in part because of rapidly changing social norms as to the roles suitable for women, in part because of the greater economic opportunities available in these fields compared to the fields traditionally reserved for women. Passage of the Equal Rights Act amendment in March of 1972 greatly facilitated breaking the barriers to women in employment situations, and made police agencies vulnerable to charges of sex discrimination. This study had a potential for major impact primarily because it was the first major study of women performing the traditional male functions in the police department, but also in that it was the first study of a sizeable number of women in any occupation which was so extreme in its masculine orientation that women were systematically denied access on the basis of their inherent "physical" unsuitability.

Other factors in addition to timeliness, however, are necessary to insure that an evaluation has the desired impact. The evaluation must be conducted with a certain degree of methodological sophistication, the data obtained must be relevant to the issues of central concern to decision-makers, there must be some means of distributing the results to those persons who are in a position to act upon them (or to publicize the data, if decision-makers are reluctant to act, in such a manner that public pressure is brought to bear), and there also needs to be a favorable climate of publicity surrounding the program and the evaluation.
Design

The design of the evaluation itself greatly enhanced its impact. The sample characteristics were critical—in particular, the fact that the sample size was large enough to permit statistical tests and to generalize from the sample studied (80 women in the first phase, 86 in the second), and that these women, unlike the women studied elsewhere, were selected, trained and deployed in a manner identical to the men in the department lent this study greater applicability. Had the women been college educated, as is required in some departments, held to different standards, or had they been assigned to other duties prior to going on patrol, this study would not have been able to present a picture of the female recruit that could be expected under normal hiring conditions, and hence the study results could be considered to be valid only in special cases, such as when the standards for females were considerably higher than for males. Other evaluations of policewomen performed prior to or concurrently with the D.C. evaluation had virtually no impact upon the judicial system, and consequently very little persuasive impact (outside of the departments in which the studies took place), since the samples were small, the duties somewhat atypical, and the women were not necessarily selected in such a manner that they could be compared with men.

The D.C. evaluation was designed in such a manner that adequate comparison groups could be obtained. Although this seems an obvious requirement, program evaluations frequently lack comparison groups (Weiss, 1970; Hatry et al., 1973; Campbell, 1970; Wholey, 1970). A simple description of the performance of the women in D.C. would have
been virtually impossible for the courts to interpret as evidence of women's capabilities, since there is no available standard of comparison for police performance. Even police departments which might have been somewhat inclined to use policewomen would have had difficulty using such data, since the variance between departments with respect to the level and type of activity as well as the ratio of police to citizens is large, and departments could not be expected to extrapolate from the D.C. Department statistics to their own departments.

In order for the study to be seriously considered by the courts and other police departments, it was necessary to conduct it in as objective manner as possible. The reasons for studying uncontaminated E and C groups were explained to members of the police department, in order to obtain their cooperation in altering the usual assignment procedures. The results were presented in terms of statistical comparisons of experimental and control or comparison groups in order to appear scientifically credible, and the evaluators refrained from advocacy positions.

The credibility of the study was further strengthened by the focus upon the issues which were posing the greatest barriers to hiring women on patrol. The evaluators spoke with a number of officers, officials and police administrators during the instrument development phase and conducted a pre-test of the Patrol Survey, from which sources the major concerns about policewomen were ascertained. Districts were selected in which the women would be put to the test. (Had the women been patrolling in a suburban department, as in one of the concurrent studies [Sherman, 1974], or in a relatively low crime area, the impact might have been diminished, since this would have failed to address one
of the most serious questions about women on patrol—their ability to handle dangerous situations.) Attempts were made to obtain as much data as possible about performance in violent situations, about arrests, injuries, protecting one's partner, and physical inadequacies (e.g., excessive sick leave, inability to pass the physical standards for agility and strength, etc.). Other purported obstacles to policewomen were investigated, such as negative community sentiment and opposition by patrolmen's wives.

The policewomen design, of course, was atypical for most program evaluations, since there was no "treatment" per se, and hence it was not program effects, but the feasibility of including another population group (women), in the regular operations of an organizational system. The results from this evaluation are particularly applicable to evaluations of the capabilities of populations which have been excluded on the basis of characteristics which are deemed to be essential for performance of the job by employers, such as race, age, various physical standards (height, strength, agility, absence of disabling physical condition, etc.), education, and so forth.

For the purpose of maximizing impact, it was necessary to obtain comprehensive data about performance, although this involved methodological costs. The researchers felt that it was more important to gather data which would address almost any possible question about women on patrol, from a variety of perspectives, which could be justly raised by police departments or the courts, than to focus on one or two carefully developed measures; although these measures might have better data on their reliability and validity, they might be of less
use to decision-makers, in that they would provide a relatively narrow picture of a complex issue. Thus, the persuasive strength of providing several indicators of performance ability and community reaction was deemed far greater than the strength of reliability and validity indices.

**Characteristics of study participants**

The Police Foundation had contacts with police departments and police associations across the nation, the Urban Institute was known for conducting research which had policy implications on a national scale, and the ICMA had ready access to city managers nationwide. The panel selected by the Police Foundation for review of the study and the report itself consisted of experts in the fields of criminal justice, law and program evaluation. Thus, the organizations responsible for the evaluation lent the study credibility with the audience for whom the study was originally intended and to whom the results were distributed: police departments, the legal profession and city managers throughout the country.

The policewomen evaluation provides a good example of the beneficial effects of the decision-makers, the program administrators, the funding agency and the research organization being relatively independent of one another. Given the controversial nature of the program, this was particularly critical, and it is not simply a function of ill will or undue personal/political interests influencing the outcome. Independence is in part helpful simply because the roles and values of the various participants differ, each with the possibility of enhancing the evaluation, and none with the capability to have all perspectives
and to excel in all roles. For example, the reluctance of the D.C. Police Department administration to assist in prosecuting or pressuring other police departments on this issue was due to the fact that this action would have been in opposition to other roles and values of police administrators, and hence it was fortunate for all parties concerned that the program administrators were not required to be involved in the utilization of the study results. Similarly, the funding agency was under no pressure to accept or reject specific evaluation proposals, nor to obtain certain results, and hence could afford to negotiate for the type of evaluation which it felt was needed, and to present the results, regardless of findings. The research organization's independence from the other agencies probably lent the study results greater credibility in the courtroom, and certainly permitted greater flexibility in determining the appropriate types of questions to be addressed, in developing the design and measurement instruments, in accommodating to changes in the program or to noted deficiencies in the early evaluation procedures, and in presenting the results than if the funding agency, the decision-makers or the program administrators had held any sanctions over the organization or the researchers. Finally, through discussion and, at times, conflict between the evaluation participants, a larger variety of perspectives were obtained than would have been possible from any one of these individuals or organizations, resulting in an evaluation effort that addressed most of the questions pertinent to interested groups in a fairly objective, relevant and comprehensive manner.
Planning for impact

The distribution of the evaluation results was planned prior to the evaluation. As mentioned above, target groups were determined, and, in the second study phase, professional editors were hired for refinement of the final report.

The role of the media in planning for impact is somewhat complex, in that: 1) the extent of media coverage is in some respects one of the impacts of the study, but 2) the type of media coverage can influence the impact upon other organizations and/or the public, and 3) the evaluators and the funding agent have limited control over the type and extent of media coverage. This control varies, to some degree, with the type and extent of media coverage allowed or encouraged in the various stages of the program as well as with the specific, newsworthy program problems or successes which occur during the program, which are generally not known in advance. That is, the program may or may not be publicized extensively prior to the evaluation results. If publicity is planned by the evaluators, and the program is publicly accessible and interesting as a news item, the evaluators lose a fair amount of control over the impact of the media upon public opinion, which may eclipse the evaluation impact. If the evaluation itself is anticipated my media personnel to be of major impact, then it has a greater chance of being publicized, and hence the media coverage itself is in part a study impact.

In the present case, the media recognized a hot topic, and Chief Wilson's announcement of the policewomen program alerted the media quickly to the issue. Extensive publicity was spurned pending the
results of the second year, primarily because of the risk of distortion due to reliance on early, tentative results, or to reporter's desires to present an interesting story, regardless of the accuracy. Interviews which were granted were carefully screened, and every persuasive effort was taken to assure that reporters presented the information obtained fairly. Prior to publication of the second year's results, however, the pressure for news was such that a near-final version was released to the press. The press release was designed to control the public impact of the study, rather than having the results leaked by more informal and less accurate sources (a conference on policewomen was to be held about one week after the press release, at which time the results of the policewomen evaluation in D.C., as well as the smaller studies in New York City and St. Louis county were to be presented and discussed).

Although there were some negative or blatantly sexist articles written about policewomen, on the whole the coverage was favorable and accurate. This was in part attributable to the fact that the department, the evaluators and the funding agency all made a conscious attempt to deal with the media in a very discrete manner, to minimize the possibly more newsworthy but potentially damaging "telling of tales," and in part to the fortunate occurrence that no major disasters occurred in the program during the two year evaluation.

The impact of this evaluation can be assessed in a number of ways, including the impact upon the legal system, the impact upon other police departments and police agencies, the impact upon the D.C. Police Department, and the impact upon further research in this area. The
total impact of this study is unknown, since this would have involved extensive survey efforts to determine. In addition, the author was not involved with the project following completion of the study. However, from the very partial impact data which are known to the author, the scope of the study's impact may be guessed.

The final report, which contained the results of the second phase of study in a form suitable for a general audience, was published in May, 1974, and the second volume, containing the methodology, tables and measurement instruments, which was designed primarily for researchers in the field, followed in December of that year. As of May, 1976, over 6000 copies of the first year's report and 13,222 copies of the final report had been distributed to police departments, government and criminal justice agencies, researchers, etc., almost half by request.

Legal

Lawsuits against a number of (primarily) urban police departments have been waged, some by the Justice Departments' Civil Right Division, others by private law firms, often as class action suits. These departments include: Detroit, Cleveland, Baltimore, Chicago, Philadelphia, San Francisco, Milwaukee, and Montgomery, Alabama. A case is currently in process against the Alabama State Police. The legal issues varied; in some cases departments were being sued for racial as well as sexual discrimination, in other instances promotional practices, height, and physical standards, such as agility, were at issue, but in all cases the suit involved the issue of hiring or promoting women in a manner similar to the men. Some of the trials were quite colorful—e.g., in Cleveland, the defense argument that shorter officers lacked the
necessary physical strength was short-circuited by the prosecution's introducing a videotape of a petite female law student (weighing about 103 pounds) with training in karate, who demonstrated, among other things, how to cope with 3 large male "attackers"—all 3 were readily indisposed—as well as by the judge (who was 5'2"), who noted that he had been a Golden Gloves boxer. In September, 1973, the federal district court in Cleveland ruled that height minimums are discriminatory, that stature aids little in providing the "leverage strength" needed for most police tasks and that physical fitness in no way pertains to height (this decision was later overturned on a technicality in Appeals Court, and a current appeal to the Supreme Court is pending).

All of the other lawsuits were settled or won out of court in favor of the prosecution. A common type of settlement was for the department to be ordered to hire between 80 and 100 policewomen. In all cases except Detroit, which occurred prior to the second phase results being available, the D.C. evaluation was utilized as a basis for disallowing sex discrimination, and testimony, usually in the form of depositions, was obtained from one of the two study authors. The plaintiff lawyers handling the case were often women, in some instances women who had themselves suffered sex discrimination in their profession. In some cases, judges and/or witnesses who had been initially opposed to policewomen were favorably influenced by the study results. Two of the cases will be described in somewhat greater detail, since they illustrate the intensity of the opposition to hiring policewomen, as well as some of the persuasive factors which played a part in these suits.
San Francisco: This case was very lengthy and complicated, and involved both racial and sexual discrimination. The department was called into question because of the almost non-existent proportions of officers of Chinese or American Indian extraction, both of which are amply represented in the city's population, and its total absence of women on patrol. During the trial, the department presented statements from about 70 officers stating that they were of American Indian extraction, which were virtually dismissed as false, and defended the lack of Chinese by indicating that they would teach non-Chinese officers the Chinese language. Four women were hired for patrol work on an "experimental" basis, and within two months (it could not be determined exactly how long they had been on patrol), the Chief produced written affidavits from all four stating that they were incompetent to handle the job of patrol (these statements were presented in defense of their not placing more women on patrol as well as in defense of their removing these four women from patrol work). The judge, who reportedly was unfavorably disposed to women on patrol, was influenced by the study results to change his mind (undoubtedly the conduct of the police department also contributed to his being less favorably inclined toward their position). A (female) member of city council, who had initially been quite opposed to the use of women on patrol, stated at the policewomen conference that she "did not want to go to the funeral of a policewoman." She later changed her mind and testified in court that women could do the job (she also was reportedly influenced by the study results). City council, which is responsible for police hiring practices in San Francisco, initially had opposed women on patrol, but now supported the idea. Peter Bloch, the project manager, testified in district court about agility testing and about the policewomen evaluation. The final result was that the judge ordered in open court that the department hire approximately 80 women, and wanted it entered on the record that women were capable of handling uniformed patrol work.

Philadelphia: The Philadelphia Police Department was completely opposed to hiring women for patrol work, and the assistant city solicitor defended their case against the Justice Department's Civil Rights Division suit. In addition to suing the department for refusal to hire women for patrol, women were attempting to receive back pay for promotional opportunities that they had been denied (for which they were ineligible as a class because of their sex). The author was deposed by the city of Philadelphia, which was representing the police department (the assistant city solicitor's office), which I will describe in some detail. Depositions are usually conducted at the place of work or residence of the individual being deposed (as was done in a case against Chicago, with the same author, using the services of a local law firm, with a lawyer and a court recorder present). Philadelphia, however, attempted to stall the case by indicating that it had not had the opportunity to depose a key witness (the author); on the suggestion of the Justice Department lawyer, I went to Philadelphia for the deposition, to avoid further delays. I was met by the assistant city solicitor and 5 or 6 other lawyers, as well as a psychologist, all of whom fed the examining lawyer (the assistant city solicitor) questions during an exceedingly
long deposition (four hours, with one five minute break). The questions were often boldly provocative, apparently in an attempt to unnerve the author sufficiently that the testimony would suffer. The entire thrust of the questioning was to discredit the study, by one of several tactics: 1) discrediting the study authors and participants, such as the observers, by indicating bias (e.g., I was asked if any of the observers were divorced, were members of feminist organizations, were lesbians, had been involved in civil rights activism, etc., and whether I had been paid to do the study, how much I had been paid, etc.); 2) discrediting the study on methodological grounds (e.g., I was asked who trained me to train the observers, what accepted research model I was using in constructing the performance observation instrument, how could I defend the various sample sizes, and, for each measurement instrument, what was its reliability index and what was its validity index (with no explanations permitted for any of these items beyond a direct response to the question); and 3) to discredit the study by attacking the inferences drawn from the data. The questions were frequently asked in a fairly hostile or derogatory tone, again with the probable aim of angering the witness to the point where testimony would be damaging to the case (the deposition is recorded by a court recorder, so none of the affect is known to the judge who reads the material). During the break, one of the lawyers for the defense began playing with another individual's crutches, pretending that they were submachine guns. The court recorder, an older black man, asked what the case was about—upon hearing that they were attempting to block the possibility of being ordered to hire women for patrol, he was incredulous. (In contrast to this deposition was the Chicago deposition, which took about an hour, was conducted in a business-like tone, and focused on asking the witness a large number of questions about the department, such as the number of officers in the patrol division, and other departmental statistics which the author would not be expected to know. It was unclear whether this was an attempt to make the author appear to be ignorant or unqualified, since it was clear soon after this line of questioning appeared that such intimate knowledge of police department functioning was not available to the author, or whether this was simply because they were not sufficiently acquainted with the study to ask the relevant questions.) Around the time of the deposition, it appeared that the city would settle out of court for a very limited number of policewomen, with the condition that they evaluate them before proceeding to more expanded programs (the explicit notion being that the evaluation might indicate that the women were unfit, and even if this failed, the evaluation would permit a great deal of stalling before actually having to comply). Philadelphia newspapers ran some "unfavorable" articles about policewomen, which focused largely on opposition by the wives of policemen. After many delays, during which time the author moved from Washington, D.C., the research secretary on the policewomen study and the statistician involved in the height study (White & Bloch, 1975) were deposed for the sole purpose of ascertaining whether the original records still existed. Both of these individuals had similar experiences during the depositions. The psychologist wrote a critique of the study, which was responded
to by Mr. Bloch (1976), who was eventually scheduled for a court appearance. (It may be relevant to note that this particular police department was studied extensively by the Pennsylvaniz Crime Commission [1974], which concluded that corruption was widespread, systematic and continuous at all levels of the department. In the opinion of Mr Bloch, departments which have widespread corruption may have greater fears of incorporating women into their ranks because women may not be as readily assimilated into the corruption system.) At the last moment, the city settled out of court. They were ordered to pay the women who had been denied promotion back-pay, and to hire 100 policewomen for patrol.

**Police departments**

The major known influence on police departments has been achieved through legal pressure. However, a survey conducted by the International Association of Chiefs of Police (IACP) revealed that, by 1973, policing was no longer the exclusive domain of men (Washington, 1974). Whereas less than a dozen women were patrolling only three years before, slightly over 1000 women were on patrol by 1973, and another 168 were holding the rank of sergeant or above. The U.S. Army Military Police Headquarters had another 1000 women "on patrol". New York City had over 400 women in such positions, another 40 to 50 departments had varying numbers of policewomen, and a large group of departments had plans to implement a policewomen program in the near future. A center was created for studying policewomen by the IACP (which no longer exists). In 1973, the Law Enforcement Assistance Administration (LEAA) formally proscribed sex discrimination, with regulations closely paralleling EEOC regulations. This was particularly significant, since LEAA contributes nearly one billion dollars annually to police departments and related agencies. The LEAA is obliged to withhold funding from qualifying agencies failing to comply. However, only

three staff members are assigned to enforcement. In 1972, 3 complaints were received, compared to 21 in 1973. In 1973, Chief Wilson lowered the height standards for all officers to five feet, New York City abolished height requirements, and a number of other departments followed suit. Departments sent representatives to view the D.C. program and/or to talk with the study authors. A center for coordinating information about policewomen was created in Washington, D.C. In one known instance, a major department which was strongly opposed to hiring policewomen created a report for use in their department which summarized all of the potentially damaging findings against policewomen. It is not known whether LEAA threatened to cut funding for departments which discriminated on the basis of sex, but there was some indication that this had happened.

Research

The D.C. evaluation has stimulated further research on policewomen, although the goals of such research were not uniformly clear. Following the initiation of the D.C. evaluation, the Police Foundation funded a study of 25 policewomen patrolling in a suburban department in St. Louis (Sherman, 1974). The results of this study, which was more modest in scope, served to supplement the larger evaluation by providing data on women in a setting which is undoubtedly less dangerous, but which may require somewhat different skills. These findings were presented at the policewomen's conference as corroborative data. Following completion of the D.C. evaluation, the Police Foundation funded another study, this one being of policewomen in other countries. In New York City, the administration changed following the massive hiring
of policewomen (Patrick Murphy, the former Chief of Police, resigned and became the president of the Police Foundation). The new administration felt that the results from the initial study of 14 women (Greenwald et al., 1974) were much more laudatory than their experience with policewomen in general. As a consequence, this study was not released for public consumption, since they strongly objected to the findings as being unrepresentative. A joint effort between the police department and a research firm with which they had worked in the past produced a proposal to evaluate the larger sample of policewomen in New York City (the design and measurement instruments in the initial proposal being modeled after the D.C. evaluation). LEAA funded this proposal, and some time later the evaluators consulted with the D.C. evaluation staff. At the present time, due to the extensive lay-offs of municipal employees in that city, the total number of policewomen has dropped sharply, to perhaps one-third or less of the anticipated sample; the effect upon the evaluation is unknown to the author.

Somewhat after the onset of the D.C. evaluation, the "leadership" ability of female police executives was studied (Price, 1974). "Height" studies which had been performed in (and usually by) various police departments were reanalyzed in a systematic fashion (White & Bloch, 1975; Swanson & Hale, 1975), since this issue pertained to policewomen as well as certain ethnic groups. The major issue with respect to height was that previous studies had concluded that shorter officers are assaulted (and injured) more than taller officers. Reanalysis did not confirm these findings.
In addition to stimulating research, the D.C. evaluation has been used in academic settings as an example of how to conduct research in the criminal justice area. The study authors were also involved in some speaking engagements which pertained to the research goals. The author was invited to speak, for two successive years, to a seminar on criminal justice, with emphasis on research techniques in the area, the project manager spoke at police training conferences and seminars, and the author, together with others involved in the evaluation and/or program, spoke to a psychological convention, which was recorded for National Public Radio. At least one department requested the results and methodology so that they could design an evaluation (and a program) within their department.

Some of the research possibilities were curtailed because of limited time and resources. For example, during the course of the evaluation, the authors were approached, but were unable, to collaborate with the revalidation of a non-discriminatory version of the Strong Vocational Interest Test by providing data about the vocational preferences of policewomen, and the author was asked to write a fairly extensive article about the methodological issues and design of performance observations of police officers. Following completion of the study, a proposed panel of researchers in the area of women in traditionally masculine vocations never came to fruition, primarily because of the relative lack of research in this area.

Media Influence

During the course of the evaluation, newspapers ran a number of articles on the program and the evaluation results, and, following
the press release of the final report, articles appeared in the Washington Post, the New York Times, and other major urban newspapers—the results being eventually reported in newspapers as distant as Australia—as well as in weekly news magazines such as Newsweek. With a few exceptions, the major newspapers were generally responsible in their reporting, particularly in the District of Columbia. This was reflected by the comments of citizens about the effect of the media upon their opinions of policewomen (the least favorable media effect was measured in the first administration of the Community Survey, just as the first women were starting patrol). The study authors as well as policewomen in the department were solicited for interviews by the newspapers, an occasional magazine, radio talk shows, and even television specials on policewomen. Not long after completion of the study, policewomen began to appear regularly on television advertisements and weekly programs, which may be the ultimate indicator of their acceptance into the fabric of society. Despite the substantial study impact, it should be noted that the proportions of women in patrol positions still remains small, and that the issue of women on patrol has not been resolved.

SUMMARY

The study was designed to respond to a variety of goals—legal, persuasive, administrative, and research. The legal objectives, which were paramount in this effort, required data about the performance of policewomen relative to the performance of policemen which could be used as evidence of whether hiring women for patrol duty would seriously undermine the effectiveness of police operations. The study addressed
the legal and other objectives by providing data about multiple performance indicators from a variety of perspectives. Due to the comprehensive nature of the data obtained, the methodological adequacy of the experimental design—for the purposes of decision-makers—the timeliness of the study, and the advance planning for impact, the results of this evaluation were widely circulated and attended to. As a consequence, the evaluation results have been used as evidence that women are capable of handling uniformed patrol in a number of lawsuits which have been levied against major police departments throughout the country. These departments have been required by the courts to hire women for patrol duty.
APPENDIX A

A QUESTION OF SEX: SHOULD MEN BE ALLOWED TO

HOLD JOBS AS PATROL OFFICERS?*

By Merry Nelle, Chief of Police

Soilington, D.C. Police Department

In 1967, the President's Commission on Law Administration and the Enforcement of Justice warned of a critical need to attract new kinds of people to police work. The Commission recommended increased hiring of college graduates, members of minority groups, skilled civilians, and men. The recommendation that men be hired for police work has been the most controversial of their proposals.

The role of men in policing traditionally has been limited and that tradition should not be treated lightly.¹ Presently, men are used for the

¹In times of national emergency men have on occasion been pressed into police service. In W.W. I when women were needed in the war production factories, men were assigned to police duties. The Soilington, D.C. Police Department is the proud possessor of a photograph taken in 1917 which shows a man in police uniform (complete with sam brown belt and gun) directing traffic in front of the White House. Of course, when his wife returned from the war, he left the police department.

*Reproduced from The Police Chiefette magazine, May 1974 issue, without permission of the International Association of Chiefettes of Police.
most part only in those few positions that obviously require a man, such as searching male prisoners and work that requires brute force. A few are now asking why a man couldn't be used for more general police jobs. They do not understand the rationale behind the traditional ways. They prefer to be guided by the questionable results of limited experiments conducted by liberals in a few cities which have tried to use men in regular patrol work.

RECRUITMENT PROBLEMS. The problems of employing great numbers of men in police work begin with recruiting. It would be so much more difficult to find suitable male applicants than female applicants. The cost of background investigations would increase significantly. It is well established that more men than women have arrest records and traffic violation records. Because of this it would be necessary to recruit and process a much larger number of male applicants in order to obtain a smaller number of candidates eligible for appointment.

Furthermore, most male applicants would probably have had military experience. This makes them suspect as police recruits. It is feared that the practice of indiscriminate violence inculcated by the military would render these men unsuitable for civilian law enforcement. It is, also, understood that persons with military background are prone to forms of expression known as "barracks humor" which would be totally

---


3Barracks humor is noted not only in vulgar talks, but, also, in vulgar graffiti. Janitors use 75% more hours cleaning bathroom walls in the men's dormitories of Penn State and Yale than in the women's dormitories at Mount Holyoke and Wellesley, a recent survey showed.
inconsistent with what our citizenry has come to expect of its professional police officers.4

MEDICAL AND PHYSICAL PROBLEMS. With automation of human work, on-the-job physical demands upon individuals are few. Yet in the job of police officer, occasional demands are made for vigorous physical performance after prolonged periods of sedentary work. The officer may spend her on-duty hours in a patrol car without any heavy exertion for two months, and then suddenly be required to run up several flights of stairs in pursuit of a robbery suspect. A woman can handle this, but such activity has been known to precipitate cardiac failure in men.

From birth to death, women endure the pressures and physical demands of work better than men. Thirty two percent more boy than girl babies die in the first week of life. As teenagers, boys suffer 95% more injuries from participation in sports than do girls. As adults, also, men are more likely than women to engage in dangerous recreational activity. Hunting brings out violent tendencies and leads to accidental shootings because large men provide good targets for other men who cannot shoot well. The temptation to regain old glories in a baseball or football game often leads to sprains and other injuries necessitating temporary release from active duty. Men are 80% more likely than women to suffer from ulcers in their middle years and 90% more likely to suffer heart attacks before retirement.

4 In addition, the strange military practice of defacing one's forearms with inappropriate tatoos, often featuring nude women would undoubtedly be offensive to current personnel.
Another physical problem, the male "time of the week" (or as Monson and Jasters refer to it, the Budweiser Bloat) caused by the excessive consumption of alcohol directly contributes to acute danger during physical confrontations. A recent Milwaukee study confirmed that 64% of male officers were unable either to see their feet when standing upright or to absorb mild abdominal blows without nausea. As Von Plexus comments, "What good are men in a fight if they can swallow better but can't take a punch?" It is asserted in reply that women, also, suffer periodic body change. But the predictable and mild nature of their monthly occurrence—which has never been demonstrated to interfere with performance of duty—does not compare to the more frequent malaise of the male "time of the week."

A final physical problem is the well-known tendency of post pubescent males to behaviorize gender insecurities through unlimited expression of facial hair. Their moustaches and beards are climatologically maladaptive, retaining heat in summer, water during precipitation, and tending to freeze during winter. Male officers claim discrimination in grooming rules and thus cause unnecessary strains of departmental discipline. Lastly, an inordinate amount of time is taken in caring for facial hair. Those men who keep it waste time primping; those who shave every day miss an average of 4.2 days per year for sick leave related to cutting themselves (a figure which rises precipitously for heavy drinkers).

UNIFORMS AND EQUIPMENT PROBLEMS. Placing male officers on regular patrol would result in considerable additional costs. It is likely that a whole new fleet of vehicles would have to be purchased since
two 6 foot tall, 250 pound men would not be able to ride comfortably for eight hours in the standard police Volkswagon now used successfully by most of the all-female departments.

Ford Motor Company conducted an experiment in San Diego and found that heavy male officers over 73 inches tall cause more damage to scout car seats than female officers. And insurance premiums went up when the insurance agent learned that accident prone males were allowed to drive San Diego scout cars during that experiment.

Finding a uniform suitable for men on patrol has been a serious and costly problem for those departments who have begun to hire men. Some say the present woman's uniform should be issued to men without change. But others object to the idea of men in skirts. Although some men in Scotland have been wearing kilts for years, skirts are not generally worn by men in this country and it is thought that the public would object to the sight of men patrolling their neighborhoods in skirts. Proponents of putting the new patrolmen in skirts say that the skirt is the most practical uniform in that it is cool in the summer and say that they do not want the men to stand out as men by wearing a uniform different from that worn by the regular female patrol officers. Nevertheless, several of the departments hiring men have ordered slacks for them to wear and in addition have issued them the traditional skirt. Women complain that men are receiving special privileges by receiving two types of uniforms. The extra cost of slacks need not be borne by police departments. If the new men don't want to wear the traditional skirt, they should find another job.
A study conducted by Susan Sunshine of the Home Economics Department of Yale University, concludes that uniform shirts worn by male officers have to be replaced twice as often as shirts worn by female officers. Factors contributing to the wear and tear on men's shirts were: too many wearings between washings, rough handling by commercial laundries, and rips and blood stains left after fights. Women's shirts lasted longer because they were changed daily, laundered at home by hand and carefully pressed, and because they were almost never torn or stained. Ms. Sunshine estimated that a department of 900 female officers and 100 male officers would have to spend over $1500 more a year on shirts alone than a department of 1000 female officers. It is important to note that although Ms. Sunshine's study had to do only with shirts, it has broad implication for all parts of the police uniform.

Cost of facilities would increase if police departments have to build separate locker rooms and rest rooms for male officers. Another additional cost would be incurred if new police call boxes had to be installed above the present ones to avoid the poor public image created by male police officers kneeling down to use the phone.

PERFORMANCE. A report just released by the City Institute gives the results of a study of the Soilington, D.C. police department, the first in the country to assign a substantial number of men to regular police duties. Policemen were involved in far more problems or incidents of serious misconduct than policewomen, ranging from traffic accidents to using a gun improperly. About 17% of the men in the study had misconduct listed in their departmental records, compared to 1% of the women. Twice as many men than women had been fired "in the best interests of the department."
The critical question is this: If we were willing to bear all the increased expense, adverse citizen reaction, and discipline problems that putting men on patrol entails, would men be successful as patrol officers? Probably not. One prominent concern—aggressiveness—has received the most attention. There is abundant research to support the notion that men are more likely to aggress overtly against others when presented with interpersonal stress or threat. For example, Bandana and Walter (1969) in a quasi-experimental laboratory study found that male recruits were six times as likely as female recruits to punch out Bobo dolls after receiving critical feedback from superiors. Hilda Toker in her book Violent Persons, reviews three decades of psychological research and concludes that men have been shown to be consistently more likely than women to act out psychopathic, homicidal, aggressive, blatant, latent, and patent impulses. In short the research shows that before men could be allowed on patrol, there would have to be costly remedial training in non-assaultive and passive-aggressive techniques for handling interpersonal confrontations.

The City Institute staffers who spent hundreds of hours observing men and women on patrol in Soilington, observed that when a prisoner resists arrest, men revert to cave-man instincts and beat the prisoner with fists or clubs. Women on the other hand rely on psychology and sophisticated modern equipment such as mace to subdue the prisoner without inflicting injury. Dr. Willie Friedwoman, of the Institute, concluded that for each one hour of male street time, 4.6 more minutes of violent police-citizen interaction would occur if men continued to patrol than now occurs per hour of female street time and that the
figure would rise exponentially over time as citizens came to expect such treatment and to reciprocate in kind. Those results would generate lost-time increases for the following categories: giving aid to stricken victims, transporting police and citizens to medical facilities, court appearances in civil suits against officers, and down time for officer disability. That cost figure of the introduction of men onto patrol could be calculated by each department with the following simplified econometric model according to Dr. Friedwoman:

\[
\frac{(n_1 + n_2)!}{(n_1)! (n_2)!} \times T^2 + p(\cos n)^3.
\]

Another controversial issue is male sexuality as it effects performance and poses a serious threat to citizen welfare. The dominance of unbridled libidos in the potential male recruit raises the very real possibility of unseemly and offensive heterosexual behavior when possessed of the attributes of official sanction. Again a finding by Bandana and her colleagues is relevant. They found (Journal of Primate Behavior, 1968) that Sailors were more likely than Waves to make passes at scantily clad, life size Bobo dolls of the opposite sex when placed conspicuously in their barracks. These tendencies could cause problems during stops and frisks, transports, and followup interviews at homes of complainants.

Patty Block evaluator of the pilot projects in Old York City and Soilington, D.C., noted that men exhibit a distinct deficiency in the

---

5Sailors were chosen for participation based on their inability to make waves.
area of observation of suspicious persons. She noted that male officers tended to keep females under observation although these persons are not generally associated with criminal activity. In particular, Block has noted that there seemed to be a direct correlation between the chest size of the suspect and the length of observations. Researchers have been unable, however, to establish any correlation between chest size and criminality. Female officers, in contrast, keep their eyes on young, athletic, crime prone males. Even if we were to assume that male officers may have stumbled on some as yet unrecognized criminal group, they strangely destroy all possible value of their observations by making peculiar whistling noises, thereby alerting the suspects.

Without dwelling on the sexuality problem too long, the point must be made that there could be embarrassments caused by allowing men to ride in patrol cars with policewomen on the midnight shift. Recently, husbands of policewomen in Old Orleans marched on police headquarters and caused so much commotion that the police department had to abandon its experiment of using men in patrol cars.

The blatant performance problems caused by male aggressiveness and sexuality just discussed must not be allowed to overshadow the subtler problems posed by men's deficiencies in social skills. The

---

6Patty Block is employed by City Institute, but funding for evaluation came from Police Foundations, Inc. a company which fears that its sales will fall off considerable if men replace policewomen in any great numbers.
City Institute's multiphasic factor analytic studies in Old York City and Soilington, D.C. proved that women are more socially adept, emotionally responsive, and gregarious than men. They possess superior skills in communication—written and oral—than men. City Institute researchers found that these were the indispensable skills needed for the 90% of police calls which are of a service nature.

MEN'S PLACE IN LAW ENFORCEMENT. The previous remarks are not intended to suggest that men have no place in law enforcement. On the contrary, men have some special qualities which are needed in police work. When riots or other uncontrollable fights break out suddenly and there is not time to gather sophisticated equipment, build barriers, or use other humane approaches to crowd control and apprehension of violent criminals, the police department needs the brute force that big men can best provide.

The question always arises, however, of what to do with the men between occurrences that require their brute strength. One suggestion is that a few can be used in patrol wagons. Wagon drivers rarely handle the routine police calls for which men are considered unsuitable and they do frequently transport male prisoners who must be searched. The City of Los Diablos has just completed a year long innovative demonstration project of assigning one man to each patrol wagon. Chief Ed Sivad reports that the project is an outstanding success! The researchers in this study used a scientific approach by employing factor analytic techniques in the construction of personality instruments which were applied to a wide range of problems in the psychology of prisoner transports. Time spent in having to call a police-
man from across the city to search a male prisoner was almost completely eliminated. The men assigned to the wagons were even effective with some of the female prisoners. Where the policewomen sometimes had to fight to get a woman into the wagon, the men could use their masculine wiles. Particularly with drunk women, the men could flirt with the woman and get her into the wagon before she knew what happened. (But hiring police officers simply for a handsome face has its limitations.)

It is, also, possible to use one man in each of the station houses which have cell blocks so that a man would be available for searches there. More than one man per station, however, might decrease the efficiency of the clerical operations that must be performed. Although it must be acknowledged that the Bell Telephone Company has begun using men as telephone operators and clerks, police departments would be wise to await the results of that experiment before hiring too many men for such work in police departments.

Other rational suggestions for greater utilization of men include traffic control where their added height makes them more visible and where they have few citizen contacts, and in police athletic programs for teenage boys where the fact that the policemen would be out of uniform reduces their tendencies to have authority and aggression problems.

IN SUMMARY, research, experience, and common sense dictate that men are not suited for the full range of police duties. It is felt that to introduce great numbers of men into police departments would seriously interfere with the efficiency of the broad police services now performed. However, a few men carefully selected can be used in
special services.

Acknowledgements. I wish to thank my many colleagues too numerous to name who assisted with this article but a special thanks goes to my husband Harry who typed the final draft. As usual, the insights here presented are all mine; and the errors, theirs.
APPENDIX B

ISSUES IN PROGRAM EVALUATION

This section will include a review of some of the more important methodological and policy-related issues in program evaluation research, many of which were illustrated by the policewomen evaluation. The paradigms for different types of evaluations, some methodological designs which are suitable for research in field settings, some commonly encountered methodological problems in evaluation research, and the factors which affect the utilization of results will be discussed.

Most program evaluations are designed to simply determine whether the program is being carried out as planned, within the allocated budget, rather than to determine the success or impact of the program. These evaluations might more properly be termed monitoring efforts, which have the purpose of increasing accountability. Such evaluations are usually descriptive, rather than statistical, and "focus primarily on the inputs and the management process rather than on whether the program is accomplishing its intended purpose (Hatry et al., 1973). There are a number of reasons for the frequency of this type of "evaluation," among them being that: 1) in many cases, there already has been a commitment to the program such that evaluation of "effectiveness" is not desirable from the policy maker's standpoint (i.e., there are no decisions to be made about the program); 2) in other instances, pilot
and demonstration programs are launched on a small scale before even the sponsors of these programs are convinced that they will be beneficial, the immediate goal being to determine whether the program is operationally feasible (Roos, 1975); and 3) evaluators vary in their understanding of the process of conducting valid program evaluations.

While these evaluations are sometimes useful to program administrators or to the agencies responsible for the program for detecting administrative problems, most writers in the field use this type of evaluation as the primary example of poor evaluation methodology. Campbell (1970), for example, takes the extreme position that descriptive evaluations are worse than no evaluations. Korchin (1975) notes that descriptions of program history and functioning may be useful in the early stages when the development of an explicit research design may be impractical. The major problem with these types of evaluations are that they may give the appearance of assessing program effects, when in fact they are simply assuming that the program is having the intended effect, without actually measuring that effect.

Evaluations which go beyond program descriptions are usually limited to measuring the program's success in reaching goals stated in the funding proposal. Most writers in the area of program evaluation suggest this as the proper method of conducting evaluations, and ignore the issues involved in identifying appropriate program goals. For example, Hatry et al. (1973) summarize the proper procedures for conducting evaluation research as follows: "1) Identify specific program objectives, 2) specify criteria for measuring progress towards these objectives, and 3) identify the population segments that are likely to
be affected by the program and on which program impact data should be provided." Weiss (1972) similarly states that the purpose of evaluation research is "to measure the effects of programs against the goals it set out to accomplish. . . (p. 4). The rationale for evaluation research (which is often inconsistent with the reality), is that it "provides evidence on which to base decisions about maintaining, institutionalizing, and expanding successful programs and modifying or abandoning unsuccessful ones" (Weiss, 1970, p. 58).

This approach basically attempts to bring laboratory methodology into field research, by pointing out basic principles of experimental design. Nevertheless, it ignores several important issues, and the measurement of such goal achievement may result in failure to address the relevant questions. In particular:

- Stated objectives are often ill-considered, dated, deliberately excessive, or irrelevant to the actual program purpose. Proposals are traditionally written in a deceptively enthusiastic style, with extravagant claims about the expected beneficial effects of the proposed program. The statements of objectives or goals are generally intended to convince the relevant funding agencies of the ameliorative benefits and the cost-effectiveness of the program or policy, rather than to portray the likely costs and benefits of such a program (Weiss, 1973). Therefore, examination of the stated program goals is likely to yield disappointing results, since the purported goals are too often unrealistic.

- Initial program proposals are generally written long before the realities of program decisions are faced by program administrators, and
even well-intentioned goals may change as a result of experience with the program. It is not possible to know in advance all of the problems inherent in the proposed goals, questions, and procedures for responding to these questions. In many instances, the initial questions will provide only a sketchy outline of the issues which should be addressed to provide a thorough evaluation of the program. Certain answers will stimulate further questions which were not apparent in the initial stages, or which might not be appropriate had the results differed. (For example, the finding that women made less than half the number of arrests made by the men was of sufficient importance that a number of other critical questions followed, including whether the women had an equal opportunity to make arrests, whether they were fairly credited for their arrests, whether the arrests held up equally well in court, and whether the arrests were necessary and desirable.)

- A program may be amply meeting the stated objectives, and yet may also be producing significant untoward effects, or conversely, a program may be utterly failing to meet such objectives but may still be performing valuable functions not mentioned as primary program goals in the statement of objectives.

Innovative solutions to social problems will probably suffer more than traditional approaches from the latter problem; thus, confining oneself to studying stated objectives will encourage conformity in strategies of reform. In addition, the extent to which a program has attained its stated goals is only narrowly informative—little information about why the results occurred, the processes which intervened between input and outcome, and the implications for improving program
effectiveness is obtained by this approach (Weiss & Rein, 1969).

Few evaluations have the flexibility to determine the actual program effects upon the population it is applied to, regardless of whether this fits the model of stated goals contained in the program proposal. Weiss (1973) suggests that this is because adequate time and resources are often not available to carefully plan a methodologically sound, comprehensive, politically and socially relevant evaluation. Nevertheless, research institutes must respond, in less than optimal proposal writing conditions, to such RFP's if they wish to survive (e.g., the policewomen evaluation proposal was written in the space of two days, prior to the hiring of two of the three research staff members). The pressure for quick results often limits the evaluator's flexibility in modifying the design to suit the actual information needs. In addition, evaluators and/or funding agents do not always have the flexibility or sufficient knowledge of the area to generate the necessary questions which need to be answered.

Fewer evaluations still adequately address the full range of policy objectives for programs (the political, economic, social and psychological costs and benefits of implementing the program), as well as the extent to which agreed-upon criteria for program success have been attained. Wholey et al. (1970) indicate that most federal programs have multiple objectives, but that there are some good reasons for failing to address these multiple objectives. In particular, the task of identifying program objectives is inherently difficult, the relationship of program activities to agency goals is poorly understood, and even agreed-upon single policy objectives are interpreted diversely by
different individuals. Weiss and Rein (1970) argue that problems with multiple objectives are inherently unsuitable for experimental evaluation, and suggest process evaluation to supplement the quantitative data. Campbell (1972) strongly disagrees with this suggestion, feeling that process evaluations can result in misleading conclusions, and suggests measuring multiple goals by using multiple indicators. Weiss (1972) further suggests, however, that evaluations are often conducted for "non-informational" reasons, i.e., that they are not proposed "in good faith," and that ignorance of the true functions of an evaluation may lead to wasted efforts, if not embroilment in political disputes.

Finally, even well-designed program evaluations are usually intended to measure the effectiveness of only one type of program strategy, without weighing the benefits of alternative strategies. This is often the case because of the expense involved in evaluating several solutions to the problem. Various solutions which have been proposed include: initiating policies on a trial basis, with the understanding that if the first one failed, other promising strategies would be attempted, with no reflection upon the administrators of the first program (Campbell, 1969); pooling evaluation funds and using the same methodology in several locations (Wholey et al., 1970); and developing standardized or compatible data systems, together with more standardized methods of data collection (Roos, 1973), making possible comparisons across discrete evaluations. It should be noted, however, that Campbell (1972) later conceded that political pressures did not permit such a rational focus on the problem, rather than the solution ("given that advocates of programs are always in competition with other persons advocating different
solutions, the posture of tentativeness seems doomed to failure . . . if adopted as a general posture, it too could be used by a political machine to explain away a recurrent ineffectiveness due in fact to corruption." [p. 31]).

All evaluations suffer from practical constraints, irrespective of the intentions and sophistication of the researchers and others involved in the effort. Some of the constraints which are common to many evaluations, and which therefore influence the design and conduct of these evaluations, are described below.

Evaluations are conducted within a limited time frame, with limited funding and populations. While a more lengthy evaluation may be desirable from the point of view of the researcher, in that it provides a better opportunity to adequately develop the evaluation design and the measurement instruments, and it may permit the development of a data base with which to compare findings after program initiation, most evaluations are reactive to immediate needs for information. Long term evaluations may be undesirable from the decision-maker's perspective, however, since 1) the delay between program initiation and program results entailed by a lengthy evaluation effort may result in the evaluation findings being of little use to decision-makers. Major program decisions may well have been made in the absence of evaluation results, and programs are often not assured of continued funding, particularly in the absence of data indicating effectiveness, since programs are vulnerable to changes in policy, priorities, and administration. Therefore, it is crucial to decision-makers to have the evaluative information available at the time when he/she must take action regarding the
program. 2) Funding agencies have limited funds for evaluative purposes, which should ideally be allocated to obtain the optimum amount of information needed about the more critical areas. The incremental value of the additional information obtained from extensive evaluation efforts should be compared with the value of more modest efforts, to determine whether the evaluation itself is cost-effective.

Study populations also suffer from practical constraints which pose, at times, serious methodological problems. Common problems include the following: 1) in most programs, individuals are not selected randomly for inclusion in the program; the population deemed most in need of the program services is likely to be the population selected (Campbell & Erlebacher, 1970). This makes selection of a control population difficult, and often the "controls" have certain advantages over the experimental group. Some of the sampling constraints are present because of administrative reasons, and thus trade-offs must occur between good program administration and good experimental design (e.g., in the policewomen evaluation, it was not possible to use male officers in the E districts as controls, since this would have overloaded these districts with "rookies," creating problems for the effective functioning of the organization).

With these issues in mind, some of the issues which arose in the design and implementation of the policewomen program, and which are applicable to many other types of field research, will be presented.
METHODOLOGICAL ISSUES

Some of the methodological issues which are particularly relevant to program evaluation research include the following:

What factors should be considered in the determination of the methodological design?

What are some of the problems in measurement which are relevant to evaluation research?

What are the advantages and disadvantages of the different varieties of measurements?

What are the issues regarding reliability, validity, and analytic techniques?

What are the issues to be considered in determining criteria for success?

Decision-makers are given the task of determining whether the program or policy under study merits expansion, continuance at the present level, or whether it is simply infeasible. The results of evaluation efforts generally have little effect upon the continuance of a program, and effects on funding levels or on the instigation of program changes are generally slight (Weiss, 1973; Buchanan et al., 1973). One of the many reasons for this is that evaluation methodologies are often not readily usable by decision-makers. Many are ill-considered and methodologically inadequate, others are exceedingly complex and technical in presentation, while still others are designed in such a manner that the results are not directly relevant to the decisions at hand. Regardless of the specific problems, the plethora of evaluation
strategies in use results in confusion for the decision-maker, who is often ill-equipped and disinclined to evaluate the relative merits of evaluations. The solution proposed earlier, of standard evaluation methodologies, would reduce the problems of interpretation and of substandard methodological design, and as a result might increase the impact of evaluation results (Wholey, 1973). A number of possible designs have been suggested for use in typical field research situations (Campbell & Stanley, 1963). Before dealing with the specific designs, however, some cautions should be taken regarding the political, methodological and practical issues which are raised by these recommendations.

In order to enforce the use of standard methodologies, it may be necessary to follow the recommendation (Wholey, 1973) of creating evaluation units or capabilities within the agencies (usually federal or state) which fund the research, and/or which are ultimately responsible for decision-making. While this proposed system of evaluation efforts has considerable merit, including comparability across evaluations, reduction of biases introduced by measurement techniques, quality control over evaluation methodologies, and increased opportunity to obtain data on the reliability and validity of measures, some questionable assumptions are inherent in this stance. Most notably, that:

1) the agency seeking the evaluation is sufficiently competent and objective to be able to determine the questions which need to be asked;
2) the agency is capable and honest enough to be able to evaluate their own projects without undue political influence; and
3) the agency will utilize the results of the evaluation as a basis for making policy decisions. This issue has been addressed indirectly by Moynihan
(1970), who notes that the executive branch of the federal government has had a virtual monopoly on evaluation results, and that "it would be almost dangerous to permit this imbalance to persist" (p. 20). The dangers cited include the disadvantage with respect to knowledge about programs which Congress and other legislative bodies necessarily labor under, the temptation by the executive branch to withhold findings which are counter to their purposes, and the pressure placed upon universities and private research agencies to produce "positive" results in order to secure contracts.

It can be further argued that all evaluations are political to some extent. The beneficial effects of having the funding agency, the evaluators and the program administrators being relatively independent of one another is particularly apparent in the case of programs of a controversial nature, in order that the generation, conduct and utilization of such research may proceed without undue influence from the political process. The virtues of independence, however, are not unmixed, and in the view of some practitioners (Wholey, 1973), the problem of researchers being too distant from the source of the decision-making, and hence designing evaluations more in accord with their own interests than geared toward providing information of use to policy makers, outweighs the concerns about political interference.

Buchanan et al. (1973) similarly argue that the quality and appropriateness of evaluation results are more often to blame for the low impact of evaluation results than "because political pressure inevitably dominates policy-making." On the face of it, however, it would seem to be a more feasible task to educate the authors and
reviewers of evaluation RFP's to clearly specify the relevant policy or program-related questions, than to renovate the entire federal system to the requisite degree of competence and objectivity in all roles, including the role of evaluator.

Certain other points about standardized methodologies deserve mention, to avoid misleading assumptions about the types of problems which such a solution is capable of addressing.

At a minimum, most evaluations have administrative, decision-making and research goals, each of which requires a somewhat different focus. Standardized methodologies, however, cannot take into account the relationship between the goals of the evaluation and the evaluation design. The determination of the types of measures which would best answer the questions which are of primary interest to decision-makers and administrators is inherently subjective, since the goals and values of the users of such information as well as the political priorities regarding the program weigh heavily in determining which types of data are relevant, how stringent the evidence of a desired effect need be, how the outcome data will be "weighted," and so forth. This is not to imply that the authors advocating standardized methodologies are unaware of this problem. Indeed, Wholey (1972; 1973) consistently stresses the need for evaluators to understand the decision-making process and the constraints or options available to the relevant policy makers or program managers. Nevertheless, it is apparent that every field of endeavor, including the hardest of the sciences, is influenced by their traditions, taboos, personal interests, and other "unscientific" factors. Advocacy of the "neutral scientist" role
tends to direct attention away from the fact that all phases of the evaluation process involve some value choices. Even the fact of performing an evaluation, as well as the types of information which are gathered, the strategies of data collection employed, the selection of the relevant outcome data, etc., are value-laden and occur within the political, psychological, economic and sociological context in which the researcher is operating. To expect the social sciences and government agencies to be free from human influence is unlikely, although the attempt to bring some order to the scientific disarray is commendable.

Irrespective of the merits of a standardized system of federal planning and implementation strategies, there is still a great need to upgrade the quality of evaluation designs, since poorly designed evaluation strategies tend to yield results which are insufficient to provide clear-cut evidence of effectiveness or lack of effectiveness of the program. Most writers in the field of program evaluation agree that the majority of current efforts are quite deficient methodologically. Even such a basic notion as the inclusion of control or comparison groups in the evaluation design is seen to be a major advance in the field. While the conduct of sound, relevant evaluations is to some extent dependent upon the competence of the evaluation staff, the availability of persons knowledgeable in the area, the practical constraints operating; the degree of politicization, and the commitment to utilization there, are specific designs which are applicable to naturalistic settings which can be used with far greater scientific credibility than the designs most frequently in use today.
DESIGNS FOR EVALUATION RESEARCH

The ideal experimental design proposed by Campbell (1969) involves randomized assignment of persons eligible for the program to the treatment conditions. This design is applicable only when the program or policy is not implemented on the entire eligible population, since this would eliminate the possibility of the control condition. Campbell feels this is by far the most superior design, since it is the most "valid" and "efficient" of the possible designs, but acknowledges the "moral" problems inherent in this type of design. These may be summarized as objections to the prototype of this design, in which the experimenter would have complete control over the program participants, who in turn would be unaware of even the fact that they were being acted upon, which is seen as authoritarian, patronizing, and undemocratic (Janousek, 1970). As will be illustrated, however, many of these problems are practical and methodological problems as well.

In order to conduct an "experimental" evaluation, it is necessary to: 1) assign persons randomly to treatments. Evans and Schiller (1970), speaking from the point of view of government decision-makers, acknowledge that randomization is the best approach technically, but object to this recommendation on practical grounds. "Our experience leads us to conclude, though reluctantly, that in the actual time-pressured and politically loaded circumstances in which social action programs inevitably arise, the instances when random assignment is practical are rare; and the nature of political and governmental processes makes it likely that this will continue to be the case.
Unfortunately, the political process is not orderly, scheduled, or rational. Crests of public and congressional support for social action programs often swell quickly and with little anticipation. Once legislation is enacted, the pressures on administrators for swift program implementation are intense. In these circumstances—which are the rule rather than the exception—pleas that the program should be implemented carefully, along the lines of a true experiment with random assignment of subjects so we can confidently evaluate the program's effectiveness, are bound to be ignored" (p. 217).

2) Enforce 100% participation in the treatment conditions (this is indicated as the ideal state for true experiments, which is only possible with captive audiences, such as prisoners). Obviously, it is impossible to avoid sample attrition in any program. Evans and Schiller (1970) note "there is likely to be nonrandom mortality from both experimental and control groups: dissatisfied members of the control population may seek to get the program services or some equivalent treatment elsewhere; the higher (and lower) ends of the SES range are more likely to move out of the area or be impossible to locate, etc." (p. 218).

3) To avoid reactive arrangements, keep the participants ignorant of the experiment and of the fact that other individuals are being given different treatments. In most programs, it is apparent to the participants that they are being treated and evaluated, even the program goals and expectations for the participants often being quite explicit. Programs are generally publicized, and the controls may be aware of the fact that they are not receiving services available to others. If the
need for program services is high in the eligible population, this increases the likelihood of the study groups becoming aware of the existence of the program. If this occurs, then "people's awareness that they are among the experimental or the controls can produce serious Hawthorne effects" (ibid).

Another basic objection to randomized assignments in real world settings is that the controls may very likely be contaminated by the experimental, since they are both drawn from the same population. They may be in the same neighborhoods, schools, places of work or recreation, or even in the same families, which "often vitiates the whole purpose of random assignment" (ibid).

In response to the difficulties encountered in conducting true experiments outside of a laboratory setting, Campbell (1969; Campbell & Stanley, 1963) has suggested several "quasi-experimental" designs, which, in his opinion, are reasonably free of methodological shortcomings, and which are particularly appropriate for evaluations of social programs.

1) The regression discontinuity design--In this design, subjects are selected from within a narrow range of ability or other eligibility criterion, and within that range, which must have sharp cut-off points, subjects are randomly assigned to E and C groups. The regression of post-test on pre-test measures is then compared for treatment groups, with discontinuity in the regression plot at the cut-off point indicating treatment effects. This design, however, is of very limited use in the evaluation of social programs, since it is applicable only to studies of individuals in the boundary zone between the needy and those
just above, or to those within a very narrow range of ability or other eligibility criterion, and not the "disadvantaged."

2) Control series design—Subjects are selected from "natural" groups of program participants and non-participants, with both E and C groups being given pre and post-tests. Campbell indicates that the traditional practice of matching groups on pre-test scores should be avoided, since this produces regression artifacts favoring the group having the higher mean score prior to the program, i.e., the control group would generally be favored in programs for the disadvantaged, which would bias against finding program effects. If such matching techniques are not employed, this design controls for those aspects of history, maturation and test-retest effects shared by both groups, since initial mean levels have been assessed. This design will falter if the E and C groups differ not only in mean level but in maturation rate, however, which may well be the case if one group has an initial disadvantage.

3) Interrupted time series design—This is the weakest methodological version of an acceptable quasi-experimental design, in which a program or policy is put into effect across the board, and hence there is no opportunity to obtain a control group. The basic idea is to compare data before and after implementation of the program, to determine changes which may be attributable to the program. In fortunate cases, implementation may be staged, permitting comparisons of early and late stage regions. In the usual case, however, the design is weak in that it fails to account for factors other than program effects which may have changed over this period of time, and which may have influenced the results, or for naturally occurring variations in
different time samples. The tendency for programs to be implemented at a time when the conditions which the program is designed to ameliorate are more acute further exacerbate the problem, since this may have been a random fluctuation from the mean which would have leveled out over time, regardless of the program intervention, or, if it represented a steady trend, essentially positive program effects might not be apparent. Without the use of comparison groups and extended time samples, however, there is no certain way of determining whether the program had an effect.

METHODOLOGICAL SOUNDNESS

Some of the major threats to the validity of program evaluation designs, many of which were illustrated by the policewomen evaluation, will be summarized.

Pre-testing: Most evaluations are not initiated far enough in advance of program implementation to permit pre-testing, and an adequate data base is frequently lacking. In these cases, the unbiased selection of control populations is necessary to permit an adequate test of the program intervention.

Criteria: The criteria for program success are often multiple, frequently lack operational definitions, may vary considerably depending on the values and priorities of the users, and are often subject to problems in interpretation. It is difficult or impossible to weight multiple criteria in a way which would be agreed upon by all of the ultimate users of the evaluation information. When there is no single agreed-upon criterion for success, the best that can be done is to
collect a wide variety of indices, which then can be interpreted by the decision-maker according to his/her particular situation. If the criteria include performance variables, then the difficulties of determining "good" and "poor" performance in real life situations are immense, since performance standards are rarely pre-ordained and quantified. Direct measures, with operationally defined criteria, are generally preferable to indirect measures, such as ratings, which tend to have problems of inconsistency, halo effects and personal bias.

Sample: 1) Group differences prior to the program may influence program results if mean differences exist and are not measured and statistically accounted for, or if mean differences are accounted for but acquisition rates differ. 2) If the E and C groups are initially selected on the basis of matched "ability" scores, and the larger populations from which the groups are drawn differ in mean level, regression artifacts (due to regression to their respective means), will occur, biasing against the treatment group which was selected from the population having the lower mean score. 3) The differential loss of respondents from the E and C groups is a common occurrence; the reasons for such loss may differ considerably for the groups, resulting in the remaining samples having systematic differences external to program effects.

Contamination: 1) The E group, the program or the evaluation may contaminate the C group (if E and C groups are not isolated in the community, program effects may spread, or reactive arrangements may occur). 2) If participants in the evaluation are aware of the treatment group to which they have been assigned, Hawthorne effects may occur.
Measurement: The issue of measures having imperfect validity, or the "irrelevant responsiveness of measures," is earmarked by Campbell (1969) as the threat to external validity which is the most relevant to social experimentation; he indicates that this "is only to be overcome by the use of multiple measures of individual imperfection" (p. 415). Since the issues in measurement affect the success of the evaluation results to decision-makers, and the techniques to maximize utilization, and most writings on the subject of evaluation do not deal with the specific problems, they will be discussed in greater detail. A summary of the major issues in measurement which were touched upon by the policewomen evaluation follows.

- The accuracy of the recording and measurement process—Variables differ in their amenability to accurate measurement, some requiring a high degree of sophistication or judgment, and hence training, on the part of the evaluator or recorder, others, such as some of the performance observations, being basically "no more difficult than counting beans" (Bloch, 1976). If program or departmental statistics are to be used, the evaluator must take care to determine whether their recording procedures in fact are reasonably accurate. (For example, in the policewomen evaluation, departmental statistics on the officer(s) assigned to a call were poor indicators of the officers actually handling the call. The dispatcher's description of an incident, which is the only recorded description unless a report is taken, varied considerably from later descriptions of the incident. Arrest statistics were incomplete, as were other departmental statistics.)
• The amenability of the variable to measurement—Events which occur with low frequency but which are critical to the evaluation may require extensive periods of observation and/or techniques to maximize the expected frequency of such events being observed or otherwise recorded. If direct observation is not feasible or desirable, and the organization or program has no efficient, functioning procedures for recording and retrieving this information, special mechanisms for obtaining such information need to be built into the evaluation at the beginning. Data which are not normally known to the organization or program staff may be difficult to obtain, regardless of evaluation staff efforts, if it is not seen as important to the functioning of the organization or program by their staff.

• Equal opportunity of treatment groups to be recorded—If performance variables are used, it is crucial to know whether the groups had an equal opportunity to perform, i.e., whether the samples were differentially available for study, whether the statistics obtained covered the same time period, and whether both groups were equally credited for the same performance. Selection and training standards, assignments, duties or other working conditions, including attitudes of coworkers or supervisors, are important to assess in determining whether the groups had an equal opportunity to perform. If differences exist, the data will be biased towards the group having the more favorable entry characteristics or work situation.

• Differential treatment of groups reflecting on measurement outcomes—Different norms or standards may apply to the E and C groups, particularly if they are in different locations or have clearly
discernable differences in personal characteristics. If the popula-
tions are not treated similarly with respect to particular variables,
this poses problems in interpretation, particularly if the variable
in question is used as an indicator of another variable.

- Personal bias affecting the measurement process—This can be
  minimized by limiting measures to those variables which can be clearly
  counted, verified, or operationally defined, and by supervising the
  administration of measures to ensure that the designated individuals are
  in fact the respondents.

- Extra-experimental variables affecting the measurement process—
  Knowledge of conditions other than the program which could have influ-
  enced the statistics obtained, and measurement of these other factors,
  when suspected, is crucial. Direct involvement of evaluators at an
  intimate level with the program and the evaluation, and personal inter-
  views of program participants may elicit some clues as to these factors,
  which might not be obvious with structured measurement procedures alone.

- Representativeness of measure—Most programs are designed to pro-
  duce complex and long-term behavioral effects, to which the evaluator
  is privy to only a very small part. Since it is impossible to observe
  a total sample of behavior, the data obtained will of necessity be
  inferential, and the task of the evaluator is to select those portions
  which are representative of typical program effects. If performance
  measures are obtained, a wide variety of behaviors may need to be
  measured. Discrete performance indicators are often subject to variable
  interpretation and the problem of determining priorities. Complex
  behavioral indices which permit measurement of interactive effects
(between performance and the context in which it occurs) and which provide a wide sample of performance indicators are preferable.

- Administrative difficulties—The collection of data may disrupt the functioning of the organization or program and produce reactive effects. If program or agency staff are required to cooperate in obtaining measures, irregularities may occur if this process is not adequately monitored, and if the evaluators lack effective enforcement capabilities. Completion rates will be dependent upon the organizational support for the program and/or evaluation, the incentives or disincentives for cooperating in the evaluation, the relevance of the questions asked, etc. If parts of the evaluation are performed by organizations other than the major contractor, there will be some costs, which are dependent upon the other agency's expertise, motivation to conduct this portion, and the degree to which the efforts can be coordinated philosophically and practically. The costs should be weighed against the political or methodological trade-offs, such as their ability to enhance impact or their greater expertise in the specific area of study.

- Credibility of measures—The credibility of the measures to the intended audience (such as administrators, decision-makers, researchers, general audiences) should be considered prior to conducting an evaluation, through consultation with representatives of these groups. The type of measure, the data collection method, the questions asked, if applicable, the instrument format and the analytic techniques should be geared towards presenting a clear, useful and comprehensive picture of the program to the relevant persons.
• Cost—The value of the information obtained from different types of measures should be weighed against the widely differing costs of different types of measurement techniques.

• Format of measurement instruments—The format and content of any instrument is dependent upon the goals of the evaluation. If the content does not address the questions which follow from these goals, the information obtained is useless. The format of different types of measurement instruments necessarily differs, as does the level of detail possible with each type of instrument. For example, observations of behavior permit the greatest degree of detail, and take into account the context in which the behavior occurs. Interviews permit considerably fewer questions, and the questions must be simply stated for most audiences. If responses are pre-coded, a certain amount of artificiality is present, but if not, coding difficulties are increased considerably. Most evaluators feel personal interviews are the most preferable, followed by telephone interviews, the least favored being paper-and-pencil surveys. Hochstim (1967), however, reports little difference in these methods except for cost, but the purpose and complexity of the interview may affect this finding. Paper-and-pencil surveys are generally fairly economical, but require formatting which may be consistently and quickly understood by respondents, which tends to exceed the difficulty of formatting interviews for trained interviewers.

• Reliability and validity of measures—The assessment of reliability and validity involves standard statistical techniques, which need not be elaborated further. In evaluations such as the policewomen evaluation, wherein no indices of this sort were obtained, rough indications
of validity were provided by the generally similar direction of the findings from several sources, a number of which were positively correlated with one another. Program evaluations frequently lack the time needed to obtain data on the reliability and validity of the measurement instruments. Since programs often have highly specific goals, it is frequently more desirable to develop new instruments which can take into account the numerous questions of the evaluation than to rely on previously developed and standardized instruments (if available), which may be less suitable for the evaluation purposes. In these cases, the construction of the instrument can greatly enhance the validity of the data obtained, i.e., questions which are comprehensible and fairly consistently interpreted by respondents, response formats which are clear and which do not "lead" the respondent, attention to the length of the questionnaire, operational definitions when possible, consultation with individuals knowledgeable about the content area, etc. If the questions are straightforward and are to be presented simply as stated in the questionnaire, rather than used as indicators of other constructs, then the necessity for reliability and validity indices is not so acute. However, it should be recognized that the methodology of studies which will be used for policy-making purposes will be closely scrutinized "even though criticism stems more from ideology than methodology" (Weiss, 1970).

RESULTS

The results of an evaluation cannot be simply described as an objective reporting of the data obtained from the measurement
instruments. A certain amount of selective judgment is required to determine the types of analyses which need to be performed on the data to respond to the relevant questions, to determine which results need further investigation, which are worthy of mention, and how to best present a clear and fair picture of the (usually) massive amounts of statistics which are obtained from the analyses. It is generally necessary to prepare summaries, conclusions and recommendations from the data, if for no other reason than to render the report accessible to the audience, many of whom will not read the report in full, others who will need a guide to direct them through the often lengthy end-product. Such summarization involves a certain degree of subjective judgment and value positions above and beyond technical expertise.

The importance of the results with respect to the questions asked is to a large extent dependent upon the political climate. Because the major goal of evaluation research is to provide information which will be useful in the decision-making process, the costs and benefits of making such decisions must be weighed against the costs and benefits of failing to make such decisions, given data with less than optimum validity and reliability. As previously mentioned, if there is strong commitment to the amelioration of certain problems, then less stringent evidence of a desired effect is necessary than in a purely academic study. In addition, since most of these programs are concerned with social problems, changes in societal conditions will alter the results obtained.

The results of a program are not simply measures of program effects. To some extent, the novelty of the program, the charisma of the
program administrator, and the presence of evaluators affects the outcome of an evaluation, above and beyond the effects attributable to the program. Even the vigor with which researchers pursue answers to the questions posed may depend upon such factors as their personal commitment to the program, the perceived uses of the data, and the degree of freedom which they are permitted in pursuing and presenting the data. In natural settings, it is not possible to eliminate the human factor, and it seems more desirable to determine ways to maximize the beneficial effects of these "non-program" effects than to eliminate them. In the policewomen program, Wilson's strong commitment to the program, together with his personal leadership qualities, greatly facilitated program success. While some might argue that these conditions are not typical, and therefore the program may not be generally feasible, such zeal, rather than neutrality is desirable for implementation of new programs.

POLITICAL CONTEXT

The political context of evaluation research has been alluded to previously. The aspects of this which should be attended to in the various phases of the evaluation process will be further explained in this section.

First, it should be recognized that "evaluation itself has a political stance. By its very nature, it makes implicit political statements about such issues as the problematic nature of some programs and the unchangeability of others, the legitimacy of program goals, the legitimacy of program strategies, the utility of strategies of
incremental reform, and even the appropriate role of the social scientist in policy and program formation" (Weiss, 1973). That is, all evaluations employ specific, usually unstated values. In assessing the effectiveness of the program in meeting certain goals, the evaluation is implicitly accepting the desirability of meeting these goals, as well as the premises underlying the program (that the issue is adequately defined, that the interventions attempted by the program are appropriate, and that the chances of program success are reasonable). Caplan & Nelson (1973) suggest that the definition of the problem (e.g., as situationally or individually based), determines the techniques of intervention as well as the evaluation strategy.

For example, the policewomen evaluation implicitly accepted the desirability of abolishing discrimination on the basis of sex, rather than on ability to perform the job, as well as the premises that the legal definitions of sex discrimination were adequate, that the program would provide an adequate test of whether discrimination was legally warranted, and that the program was not an unreasonable venture, i.e., that women stood a reasonable chance of performing capably.

Second, evaluation RFP's are by no means uniform stimuli to respond to, and hence require some active decision-making by evaluators. RFP's differ widely as to clarity, relevance to the issues at hand, hidden agendas, and politicization (political clout of the evaluation, political pressure to look at or obtain certain results, political influence in the determination of contract awards, etc. Evaluators similarly differ in the degree to which they are willing to undertake evaluations when the questions are poorly formulated, which may require
selective blindness toward some issues, and in their tolerance for controversy, which usually envelops research relevant to social policy decisions.

The extent to which the goals and values of decision-makers need to be addressed in determining the appropriate questions to be asked and the evaluation strategies has been discussed previously. Since program evaluation is designed to be useful for policy decisions, it is apparent that there is a close relationship between policy-related and methodological concerns. It may be useful at this point to indicate some of the major ways in which the policewomen evaluation was influenced by the "political" process. First, the evaluators took care to ascertain the questions which were relevant to police administrators. Second, the variety of measurement instruments which were utilized served to provide a variety of perspectives, to provide back-up when a particular measure was lacking, and to provide some evidence of concurrent validity. Third, the design of the evaluation, the measurement instruments and the analytic techniques, together with endorsement by experts in the field, lent the study some scientific credibility. Finally, in the analysis and interpretation of the data, data which were viewed as more compelling or more critical to decision-makers were given more attention than other findings which were likely to be less influential in determining policy. For example, since some police administrators were likely to view sex differences in arrests as an important and damaging finding, a great deal of care was taken to explore the reasons for this finding.
The evaluator has some choice regarding the extent to which he/she becomes involved in interpreting the policy implications. If evaluators do not question in advance the likely uses of their evaluation efforts, the results of even the most "objective" evaluations are potentially corruptible to serve the ends of those for whom this information was obtained.

**IMPACT**

The results of a program evaluation (favorable or unfavorable), are not the only factor or even the most important factor in determining whether or not the program should be continued, increased or decreased in scope and funding, modified in purpose, methods, administration, etc., or discontinued (Weiss, 1973). Other factors, including budgetary constraints (or availability of funding for such purposes), political priorities regarding programs (including Congressional, departmental, legislative or other interests and requirements of the moment, pressure from consumer advocacy groups as well as opposition lobbying or pressure, public sentiment, media influence, etc.) may far outweigh the effects of even the most valid, reliable evaluations.

Weiss (1973) feels that successful evaluations, that is, evaluations which are well-received and responded to, are generally those in which the researcher accepts the values of the decision-makers. In the policewomen evaluation, the values of police administrators and of the legal system had to be addressed. For example, the implicit assumption of the necessity for police officers to enforce the laws and to keep order was not challenged, nor would the study have been
received as a legitimate enterprise had it been. Certain cherished traditions were not challenged, even in the discussion of results, such as the assumption that all police officers must undergo uniformed patrol duty prior to being accepted in a specialized position. The performance criteria utilized in the study held all officers under study responsible for performance on routine patrol, rather than allowing alternative skills to receive equal weight. Certain performance criteria which are generally held in high esteem by police departments were questioned (e.g., arrests), but caution was taken to focus on the methodological weaknesses of these criteria, or to appeal to other existing values within the system (e.g., effectiveness with the community). The major challenge to the traditions and values of police administrators was virtually confined to the study issue of whether women were capable of handling the job. Had there not been strong sentiments which favored this challenge within the legal system, together with some dissention in the ranks of police administrators, it is unlikely that the evaluation would have succeeded in having the impact which it enjoyed. Thus, although the program was tested as if it were independent of the social and institutional structures within which the issues were identified and treated, this in fact was not the case.

The specific factors which contributed to the impact of the police-women evaluation, and which in many cases are more generally applicable, include the following:

1) The political and social timeliness of the study—Evaluations of programs which are relatively new, and about which there has been some controversy, stand a greater chance of having a major impact,
since the results are needed for decisions which are imminent and public. In such cases, it is essential for the evaluation to conform to the time schedule of the program decisions or to at least have reached completion while the program is still an important issue to decision-makers, the public, and other groups which could have influence upon the utilisation of the study results. Many evaluations, however, are not so fortunate, since they are basically monitoring a policy decision which has already been made, or which is not likely to change as a result of evaluation efforts. Such evaluations cannot be considered in the same light as those intended for use in policy-making.

2) The design and other methodological characteristics of the study—The samples studied must be clearly defined, sizeable enough and sufficiently representative of the populations about which the study is to be generalized to permit generalizations beyond the particular program. The program characteristics should be typical and repeatable within the framework of existing or feasible future budget expenditures. The evaluation design should permit demonstration of program effects (or, in the case of the policewomen evaluation, demonstration of similarities and differences of the population group). The study should be sufficiently comprehensive to respond to the major questions asked by decision-makers, and attentive to the multiple criteria which may be operating. These questions and criteria are generally not delineated for the evaluators, in which case it is necessary to be particularly sensitive to the perspectives of the users of such information, to be aware of the drawbacks of the various criteria and measures, and to
collect a wide variety of data by different methods which would bear on these questions.

3) Evaluators and/or funding agents should consciously plan for maximizing the impact of the study prior to completion of the study, including planning for the distribution of results to relevant groups, consultation with program administrators and decision-makers during the planning, implementation, analysis and interpretation stages, to increase the likelihood that the results will be anticipated as useful sources of information, and planning for the role of the media, to the extent that this is possible and applicable.
APPENDIX C

Descriptions of Samples, Instruments and Analytic Techniques by Type of Measurement Instrument

Appendix C provides a description of the measurement instruments as well as a summary of the populations and analytic techniques utilized with each instrument in the D.C. policewomen evaluation. Major changes which occurred between 1972 and 1973 are indicated. Unless otherwise noted, the significance of group differences was determined by tests of mean differences (t-tests) in 1972, and by tests of distribution differences (chi-square) in 1973. For a description of the rationale, difficulties in administration, and issues in interpretation of these measures, refer to the Methodology section. A more detailed description of the samples and analytic techniques can be found in the volumes containing the study methodology for each year of this study. The measurement instruments utilized in 1973, which are identical or highly similar in most respects to the instruments used in 1972, can be found in Appendix D.

---

1 This section is adapted from Bloch and Anderson, 1974b, pp. 1-47, and Bloch and Anderson, 1973, pp. 1-47.

2 Ibid.
The Departmental Rating Form is a standard departmental form which is completed for all officers at the close of their probationary year. These ratings are part of the officers' permanent record, and form part of the basis for salary and promotional determination.

**Instrument Description**

The rating categories on this form included: "bearing and behavior," "human relations," "learning ability," "knowledge and skill," "acceptance of responsibility," "written expression," "oral expression," "performance of duty," and "overall evaluation." The scaled response categories, as illustrated in Figure 1 had pictorially suggested proportions of officers who should fall into each category.

<table>
<thead>
<tr>
<th>unsatisfactory</th>
<th>marginal</th>
<th>average</th>
<th>effective &amp; competent</th>
<th>excellent</th>
<th>exceptional</th>
<th>outstanding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Sample**

The Departmental Rating Form is filed in the officer’s personnel jacket. From the total sample of 86 women and 86 men, personnel jackets were located for 84 women and 82 men. Performance ratings had been completed on 71 women and 71 men, the remaining officers having been on the force for less than a year. However, the sample on which the analysis of this data was performed consisted of only 66 women and 68
men, owing to the extensive delay in locating 9 of the jackets.

Analysis

Since these forms were completed only for officers who had been on the force for a year's time, these ratings were only available during the second year of study. The partial sample of 66 newly hired women and 68 comparison men were compared on each rating category. The total sample of available ratings (71 women and 71 men), however, was utilized in the correlational study of the relationship between background and performance (Super-Personnel File).

THE CHIEF'S SURVEY

The Chief's Survey was initiated by Chief Wilson to assist him in determining whether women could be used interchangeably with men. The Urban Institute assisted in the design and content of this instrument. The survey requested information on assignments and performance variables, and also contained an officer rating form. It was distributed to district commanders, who were requested to have these forms completed for all of the designated officers, regardless of whether they had resigned, had transferred to another district or assignment within the district, or were still in their original assignment. The survey was usually completed by the officer's squad sergeant (squad sergeants usually supervise about 10 officers).

Instrument Description

The Chief's Survey requested information about: 1) daily assignments of these officers over a 6 month period, 2) performance statistics for the duration of the officer's assignment to the district, such
as the total number of arrests, letters of commendation or complaint, job-related injuries, etc., and 3) ratings of the officer on eight patrol skills. Officers were rated on the following criteria: "dealing with the public," "handling a disorderly male," "handling a public fight," "handling an auto accident involving an injury," "making a crime report," "handling a disorderly female," "protecting a partner from violence," and "in general, performing street patrol."

Response categories consisted of lines which were unmarked except for the extremes, which were labelled "extremely poor," "average," and "extremely good," with the caveat "considering length of street experience" after each label. Ratings were to be made by placing a mark on the line. Positions on the lines were later converted to numerical values ranging from 0 to 8.

**Sample**

In 1972, surveys were requested for all of the newly hired women and comparison men, all of whom were assigned to the E and C districts, and a partial sample of (22) reassigned women, consisting of the reassigned women who were assigned to the E or C districts. At the time of the survey, most of these officers were on regular patrol duty. Completion rates for the new women, comparison men, and reassigned women were 100, 86, and 77 percent respectively, the overall completion rate being 91% (see Table 10).

In 1973, surveys were requested for the total sample of 86 new women, 86 comparison men and 34 reassigned women, who by this time were assigned to all of the seven police districts. Overall completion rates were about 80%, which was somewhat lower than in the first
phase, but a strict comparison between the two time periods is not possible, since 1) if officers were assigned to more than one district or division during the designated time period, and only one rating was received for the officer, this was considered a completion; and 2) some officers had resigned by this time, and others had transferred to central headquarters or training academy, to which surveys were not sent. Recall of the performance skills of such officers would be expected to be poorer than for officers with whom they currently worked. When only the officers who were still on regular patrol duties were considered, completion rates for the two years were quite similar. Surveys were completed for 100% of the new women and 76% of the comparison men on patrol in 1973, as compared with 100% and 86% respectively in 1972.

In the event that an officer was rated by more than one official, the ratings which were used in the analysis were selected from the district or assignment within the district in which the officer had spent the greater number of total working days in the specified six month time period. There were seven such duplicate ratings, all of which were for policewomen. The ratings which were not used in the analysis often were incomplete (an average of 4.9 out of 8 ratings per officer were completed on the deleted duplicates, compared with 7.5 out of 8 ratings completed on the sample used in the analysis), which suggests that the deleted ratings had often been completed by officials who were unfamiliar with the officer's performance.
Seventeen ratings were suspected to have been completed by the officers themselves, rather than by their supervisors. These ratings were analyzed separately from the "valid" ratings (refer to pp. 62-63 for a more detailed discussion of this issue).

Table 10: Chief's Survey Sample

<table>
<thead>
<tr>
<th>Officers</th>
<th>1972 Distributed</th>
<th>1972 Returned</th>
<th>1973 Distributed</th>
<th>1973 Returned</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Women</td>
<td>80</td>
<td>80</td>
<td>86</td>
<td>71</td>
</tr>
<tr>
<td>Comparison Men</td>
<td>80</td>
<td>69</td>
<td>86</td>
<td>54</td>
</tr>
<tr>
<td>Reassigned Women</td>
<td>22</td>
<td>17</td>
<td>34</td>
<td>30</td>
</tr>
</tbody>
</table>

*Nine ratings of new women and 8 ratings of comparison men were deleted from the analysis because it was suspected that officers had completed these ratings themselves. Data on assignments, injuries, etc. were accepted, even when the officers were suspected of completing their own forms.*

Analysis

The major comparisons made on this survey were between the new women and comparison men. In addition, reassigned women were compared with new women and comparison men, although there was clearly no way to adequately match these women with any group of officers. In the second phase of study, sex by assignment comparisons were also made, the two assignment groups being officers on regular patrol (officers assigned to scout car or foot beats at least 70% of their working days between January 1, 1973 and June 30, 1973) and officers in the station (officers spending more than 30% of their time performing inside assignments). This breakdown was necessary to evaluate women on
patrol, since by this time sizeable numbers of women were no longer assigned to patrol duties. Racial differences between comparison men (but not women, due to the small sample of white women) were also tested.

OFFICIAL'S SURVEY

The Official's Survey contains both an attitude survey designed to measure police officials' attitudes toward policewomen and a performance rating form, on which officials were to rate the new women, reassigned women, and/or comparison men they were familiar with in their district. Officials holding the rank of sergeant, lieutenant or captain in the E and C districts were surveyed.

Instrument Description

The attitude survey was an adaptation of the Patrol Survey for officials. Attitudes toward policewomen and the impact of policewomen on the department and attitudes toward the job were elicited. A separate insert, which varied by district, contained a list of the names of all of the new women, comparison men and reassigned women who were currently or who had been at one time assigned to the particular district. In 1972, men in the E districts who had joined the force during the period when the new women were hired were also included on these lists. Officials were asked to rate only those officers whom they knew well enough to form an informed opinion. Thus, it was possible for an officer to be rated by no officials, by more than one official within the district, or, if they had transferred to another district, to be rated in both districts. Officers were rated on the
following criteria: handling "domestic fights," "street situations involving potential violence," "upset or injured people," and being "a generally competent patrol officer." Ratings were to be indicated on lines in which the midpoint was labelled "average" and the extremes "unsatisfactory" and "excellent." Marks on the line were later transformed to scale values of 0 through 8. All ratings were anonymous, the forms being mailed directly to the Urban Institute rather than being channeled through the police department.

Sample

In 1972, surveys were distributed to the E and C districts, with instruction that they were to be completed by all captains and lieutenants and two-thirds of the sergeants in these districts. In the second phase of study, surveys were to be distributed to all sergeants, captains and lieutenants in the E and C districts. The actual number of surveys returned in the two periods was similar (89 in 1972, 84 in 1973), but the return rate was lower in the second year of study, since a larger number of officials were sampled during this phase. Return rates were only approximations, however, since it was not

---

3 Information from police department sources indicated that there were 3–4 captains, 7 lieutenants and 30 sergeants per district who were in the patrol division (a relatively small number of additional officials of this rank are assigned to other divisions). Thus, the number of surveys sent to each district corresponded to these figures (the 2/3 sample of sergeants sampled equalling 20 sergeants per district.) Based on an estimated 30 officers per district sampled, the return rate for 1972 was 73%. Captains and lieutenants had a higher return rate than sergeants.
possible to determine the exact number of eligible officials (see Table 11).

Table 11: Officials' Survey Ratings Sample

<table>
<thead>
<tr>
<th></th>
<th>1972</th>
<th>1973</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Eligible Officials</td>
<td>Officers Rated</td>
</tr>
<tr>
<td>New Woman</td>
<td>80</td>
<td>67</td>
</tr>
<tr>
<td>Comparison Man</td>
<td>80</td>
<td>62</td>
</tr>
<tr>
<td>Reassigned Woman</td>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td>Men in E Districts</td>
<td>--</td>
<td>35</td>
</tr>
</tbody>
</table>

Multiple ratings of officers were possible. In 1972, 89 officials completed these ratings. In 1973, 84 officials completed the Official's Survey (attitude survey), but only 76 completed the ratings. In 1972, surveys were distributed to approximately 122 officials, and in 1973, to approximately 162.

During the second phase of study, the researchers obtained the D.C. Police Department annual fiscal report for FY 1972, which indicated the number of officials of all types per district for that year. These figures were somewhat at variance with the baseline which the researchers had been using for comparisons of return rates in 1972 (see ff#3). The fiscal report indicated the numbers of officials per district in 1972 were as follows: E districts (ID-59, 7D-52), C districts (5D-62, 6D-43). Some of these officials were in divisions other than the Patrol Division, and hence would not have been expected to have completed these forms, but the exact number could not be determined. The districts varied widely in the number of additional surveys requested, for no apparent reason. Therefore, the return rates reported for both phases of study are an approximation based on the earlier estimates of 40 officials holding the rank of sergeant, lieutenant or captain per district who were in the Patrol Division, rather than an absolute return rate based on the number of surveys actually received by appropriate officials. The anonymous nature of the survey precluded further investigation of completion rates.
Analysis

The attitudes of officials were compared by rank (sergeants vs. captains and lieutenants), race and district (E vs. C district officials were compared only on the "patrol situations" and "traits" questions). In addition, officials were compared with male officers on questions common to both the Patrol and Official's Surveys, and officials were compared over the two time periods to determine whether attitudes had changed as a function of experience with the policewomen. On the officer ratings, the major comparison was between new women and comparison men, with secondary comparisons between reassigned women and new women, and reassigned women and comparison men. In 1972, ratings of men in the E and C districts were also compared. (In the event that an officer was rated by more than one official, the average rating was obtained for that officer prior to comparisons by sex.) Ratings were further analyzed by rank of the rating official. Ratings given by sergeants were compared with ratings given by captains and lieutenants, the latter two being combined because of their small sample size. In the second phase of study, a sizeable number of officers were no longer on regular patrol duty, and hence additional analyses were performed to determine whether officers differed by assignment as well as by sex. New women and comparison men were grouped according to the types of assignments which they were performing as of July, 1973, including: "patrol" (foot beats and scout car), "other patrol" (scooter, wagon, tactical squad, morals), "inside" (station and other assignments whose primary duties do not involve street duty), and "resigned or terminated." The small sample of male
and female officers who had resigned or who were on "other patrol" assignments, and of men on "inside" assignments precluded statistical comparisons of any assignment groups other than men and women on patrol assignments. Ratings of male and female officers in the two time periods were also compared by rank of the rating official.

STRUCTURED OBSERVATION

Trained observers rode with the newly hired women and comparison men while they were on scout car patrol. Observers recorded detailed data about all incidents which occurred during the course of each tour observed. Observations were conducted in scout cars only, due to the difficulty of observing other types of duties and the different nature of the incidents usually encountered on other patrol assignments.

Instrument Description

The Structured Observation instrument was used to measure officer performance and citizen satisfaction with police actions. The Structured Observation instrument consisted of a "Profile Sheet," "Incident Sheets" and "Brief Incident Sheets," all of which sought information about the officer under study and his/her partner. The Profile Sheet was completed for every tour and contained questions on the background of the officers, the number of times the officer had patrolled with a woman, accidents and near-accidents during the tour, and, if the officer was patrolling with a partner, ratings of the officers' working relationship.
An Incident Sheet or a Brief. Incident Sheet was completed for every incident encountered on the tour. On the Incident Sheet there were numerous pre-coded descriptions of officer and citizen actions, characteristics and attitudes, as well as the context of the incident. The Brief Incident Form, which simply described the nature of the incident and the major action taken by the officer(s), was used only when the incident was relatively minor in nature, short in duration and when no strong citizen reactions occurred.

Sample

All of the new women and comparison men assigned to patrol duties were eligible for observation. In 1972, officers were not observed until they had a minimum of two months street experience. Observers rode with these officers on day and evening shift tours (excluding midnight shifts) during the period from September through December of 1972, for a total of 199 tours (8 hour shifts). Of these, women were observed for 93 tours and comparison men for 106 tours. Slightly over 600 incidents were observed.

In the second study phase, observations were conducted primarily during July and August of 1973. During the observation period, approximately 48 of the original sample of (86) women and 62 of the (86) comparison men were on regular patrol assignments. Some of these officers were frequently given daily assignments to the station, wagon, foot beats, etc., where they could not be observed, with the result being that only 33 women and 48 men were observed at least once during the two month period. Only evening shift tours were observed, weekend observations were emphasized, and Sundays were eliminated for civilian
observers in an attempt to increase the number of incidents observed. A total of 191 tours were observed, 45% of which were tours in which women were observed. Approximately 960 incidents were observed, of which slightly over half were dispatched incidents. Of these, about one-third were incidents in which the police left the scene without performing any activity (wrong address, nothing found, police were not wanted, or other police handled situation). All but six of the remaining incidents were officer-initiated, the most common types being talks with citizens and giving traffic tickets or warnings (see Table 12).

<table>
<thead>
<tr>
<th>TYPE OF POLICE UNIT</th>
<th>NUMBER OF TOURS OBSERVED</th>
<th>DISPATCHED INCIDENTS/TOUR</th>
<th>OFFICER-INITIATED INCIDENTS/TOUR</th>
<th>TOTAL INCIDENTS/TOUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Man</td>
<td>10</td>
<td>2.10</td>
<td>3.20</td>
<td>5.30</td>
</tr>
<tr>
<td>One Woman</td>
<td>42</td>
<td>3.17</td>
<td>*</td>
<td>4.22</td>
</tr>
<tr>
<td>Two Men</td>
<td>94</td>
<td>2.56</td>
<td>2.72</td>
<td>5.28</td>
</tr>
<tr>
<td>One Woman-One Man</td>
<td>43</td>
<td>2.45</td>
<td>2.14</td>
<td>4.59</td>
</tr>
</tbody>
</table>

*a Includes incidents in which there was nothing found when the police arrived.
*b Excluding six incidents initiated by citizens on the street.
* Significant difference.
+ Tends to show significant difference.

In 1972, the ICMA selected, trained and supervised the observers. No police observers were used and both blacks and males were underrepresented (68% of all observers were female, and 73% of all observers were white). In 1973, the proportions of observers were approximately equal by race and sex. Female police observers were recruited by sending letters to the reassigned women (who were not under observation), and announcements were made within the districts for male
volunteers. All police observers had more than a year's seniority. Twelve police observers were recruited and trained; six completed at least one observation tour, and only four completed more than one tour. A total of 27 civilian observers completed at least one tour; civilian observers did not have as high an attrition rate, and it was easier to hire replacements for them.

Observers were rotated between the E and C districts to minimize the effects of observer bias. Observers attended four training sessions, each of approximately two hours duration, as well as an eight hour "training ride."

A training manual (see Bloch & Anderson, 1972, pp. 226-246) was written to acquaint observers with police terminology and procedures and to explain all terms on the Structured Observation instrument. The training sessions consisted of discussions about the instruments and the training manual, completion of the Structured Observation incident forms after viewing police films of various types of incidents, and discussion of difficulties and inconsistencies in completing incident forms for both filmed incidents and the incidents encountered on the training ride. Each item on the incident form was discussed in detail to achieve a relatively uniform understanding of terms.

The officers observed were selected by a random assignment procedure among the pool of matched officers, with a maximum of five tours

---

5Films were obtained from Documentary Educational Resources, Center for Documentary Anthropology, Somerville, Mass. The films consisted of actual police incidents which occurred in Pittsburgh, Pa.
per officer.

Analysis

The major types of comparisons were between incidents handled by men patrolling alone and by women patrolling alone, and by men patrolling with a (male) partner and women patrolling with a (male) partner. Two tours were deleted from these comparisons—in one, a comparison man was patrolling with a woman, and in the other two females were patrolling together. Key variables were selected for analysis, including: the number and type of incidents; officer-partner interactions; backup by other officers; citizen and officer attitudes and level of respect; types of citizens encountered in incidents (e.g., angry, drunk, etc.); and average citizen reactions to "conversation," "advice," "action" and "force" when used by the officer or partner (a number of different officer actions were included under each of these four categories; e.g., "advice" included "advised and settled," "assisted citizen," "took report," "voluntary transport," etc.). When more than one citizen per incident reacted to any category of officer action (e.g., "conversation") the reactions of all citizens were averaged for that incident. Reactions of citizens were analyzed by various characteristics of the citizens as well, such as age, sex, emotional state, and their relationship to the situation (e.g., victim, complainant), as well as the type of situation (e.g., robbery, domestic fight), to determine whether these characteristics affected their reaction to the officers.
The number, type and outcome of incidents observed by male and female observers was compared by sex of officer to determine whether observer bias had occurred or whether the observer's sex had affected the officer's type or level of activity.

SERVICE SURVEY

Citizens who were directly involved in incidents which were handled by the new women or comparison men were later interviewed about this experience. Respondents were asked to describe the events which occurred in the incident, their perceptions of and level of satisfaction with the way the officer(s) handled the incident, and their general attitudes toward policewomen.

Instrument Description

The Service Survey was an interview in which citizens were asked to describe the incident and their reactions to the officers who handled it. When the officer under study was patrolling with a partner, citizens were asked about their reactions to the "team" of officers, since it would not have been possible to distinguish the officer under study from his or her partner when the officers were of the same race and sex. When a female officer was patrolling with a (male) partner, a few questions were asked about her performance and the leadership pattern.

The screening form for the Service Survey contained several questions pertaining to proper identification of the incident and of the officers under study. Separate forms were used for ease of interviewing respondents whose call had been answered by officers patrolling
alone vs. officers patrolling with a partner. The substance of the information sought on each form was basically the same, except that two questions about the interaction between the officer and his or her partner were deleted when the officer patrolled alone.

The content of the Service Survey included a number of questions which were as similar as possible to the questions on the Structured Observation form, including whether anyone present was sick, injured, dangerous, angry or upset before or after the police arrived, ratings of the officers' handling of the situation, questions regarding force, threats, insults or arrests made by police, citizen reactions to the specific officer actions, and ratings of the officers' general performance, respectfulness, understanding of feelings and comprehension of the situation. Respondents also were asked several questions from the Community Survey, including their preference for an all-male or mixed sex team, whether they expected policewomen to be more or less respectful than policemen, what effect they felt policewomen would have on crime, how they felt the community would treat policewomen compared to policemen, and whether they approved of women on patrol and women's rights. When a policewoman was present, citizens were asked to rate her performance and to indicate whether their experience had affected their attitudes toward policewomen.

Sample

In the first phase of study, all police calls which were handled by new women or comparison men over a six and one-half week period (September to October, 1972) were reviewed. Midnight tours were eliminated, as were calls which had been handled by officers having
less than 2 months experience on patrol duty and calls of a minor
nature. A sample of citizens who had been present during any signifi­
cant police-citizen interaction was then obtained from the remaining
calls, and Chilton Research Services then attempted to interview these
citizens by telephone or in person. This method had a number of draw­
backs, most notably that it could not be ascertained whether an inci­
dent had in fact been handled by the officers under study. Approxi­
mately half of the citizens interviewed could not correctly recall the
race and sex or the officers involved, which led the evaluators to
believe that police records were not a reliable indicator of the
officers who had handled the incident. In addition, many records were
illegible or did not contain sufficient information to contact any
citizens involved, a number of interviews could not be completed due
to respondents disclaiming knowledge of the incident or refusing the
interview, and certain types of incidents were disproportionately
sampled ("drunk" and "sick" cases were undersampled because the complain­
ant could not be contacted or was unaware of the police contact, and
traffic accidents were over-represented, because they require a
police report, which has more accurate and detailed information about
the participants' whereabouts, and perhaps because citizens involved
in traffic accidents are more likely to have stable residences and
telephones than citizens involved in certain other types of incidents.)

During the second phase, this problem was avoided by using only
incidents which had been observed in the Structured Observation pro­
gram, to ensure that the officer under study was present. Only inter­
views in which the officer(s) had been correctly identified were
reported. The results from the Structured Observations in 1972 indicated that the sex of the observer did not affect the type or frequency of dispatched incidents or the officer or citizen reactions in these incidents, nor did the observer's presence differentially affect male and female officers. Possible observer effects were considered less serious than the problems associated with drawing the sample by using police data about the people served by police. Only one citizen was interviewed per incident, and minor incidents (noted on the Brief Incident Form) were not included.

In 1972, 507 interviews were completed. Female officers (and their partners) handled 36% of the incidents for which calls were completed, comparison males handling the remaining 64%. In all but 4 incidents (.8% of the calls), the officer under study was patrolling with a partner. In 1973, 376 incidents which could qualify for Service Survey interviews were observed. Names of citizens were obtained in 241 incidents, and interviews were completed for 135 incidents, of which 43% were handled by female officers. About one-quarter of the interviews conducted in both time periods were conducted in person, the remainder being by telephone. In an attempt to increase the completion rate, letters which asked respondents to call Chilton were sent to individuals for whom no telephone number was listed, or who could not be reached in four telephone attempts. Few respondents called Chilton, but this method was of some help in that letters were returned for eleven respondents who no longer lived at the given address.
The typical interview was about the performance of two male officers or of one female officer patrolling alone. Of the completed interviews, 54% were handled by females patrolling alone, compared with 15% by males patrolling alone. Only data from interviews in which the officer was correctly identified were reported, however, which changes these figures to 64% of the women and 16% of the men who were patrolling alone in these incidents. Backup officers arrived in 21% and 30% of the reported incidents handled by female and male officers patrolling alone, respectively. The number and sex of the respondents interviewed in 1973, by type of police team that had handled the incident, are illustrated in Table 13.

Table 13: Service Survey Sample Characteristics (1973)

<table>
<thead>
<tr>
<th>Respondents</th>
<th>Male</th>
<th>Female</th>
<th>Total N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males with Partners</td>
<td>50%</td>
<td>50%</td>
<td>54</td>
</tr>
<tr>
<td>Females with Partners</td>
<td>50%</td>
<td>50%</td>
<td>16</td>
</tr>
<tr>
<td>Males Alone</td>
<td>60%</td>
<td>40%</td>
<td>10</td>
</tr>
<tr>
<td>Females Alone</td>
<td>62%</td>
<td>38%</td>
<td>29</td>
</tr>
<tr>
<td>All Males</td>
<td>52%</td>
<td>48%</td>
<td>64</td>
</tr>
<tr>
<td>All Females</td>
<td>58%</td>
<td>42%</td>
<td>45</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td></td>
<td>109a</td>
</tr>
</tbody>
</table>

*aInterviews in which the officer(s) was not correctly identified (N=25) are not included.
Analysis

In 1972, telephone and personal interviews were first analyzed separately and then combined, since the results were very similar. Incidents in which officers were correctly and incorrectly identified were analyzed and reported separately (combined data was also reported). Within the correct and incorrect identification categories, questions were analyzed by race and sex of officers and respondents, as well as by age of respondent. Respondents were also grouped with respect to whether someone present was 1) dangerous, 2) angry, 3) upset, or 4) none of these, with comparisons made within these groups of teams with male vs. female officers. The specific comparisons included whether or not women used more or less force than men in similar situations, whether they calmed or aggravated situations by their presence, whether their presence affected the frequency of arrests, threats or use of force by male officers, whether the community reacted more or less favorably to policewomen, whether citizens differed by race, sex or age in their reactions to policewomen and whether the race or sex of the officer affected citizen reactions.

The Service Survey differed from many other surveys in that it contained a large number of independent variables which could have affected the officer ratings, and which were not necessarily distributed proportionately by sex of officer, such as: the characteristics of the respondents, the interview characteristics (telephone or in person); whether the officer was patrolling alone or with a partner; the characteristics of the officer and his/her partner; and the characteristics of the incident itself (whether any citizens were
angry, etc., and the specific actions taken by officers).

In 1973, three correlational matrices were computed in order to determine which variables were most influential. Background variables (including the respondent's age, race, sex, income and relationship to the incident [complainant, victim, accused or participant/informant], the officer's sex and race, whether the officer was patrolling alone or with a partner, how many officers arrived and whether the interview was by telephone or in person), were correlated with each other to determine whether any of these variables were so highly related to each other that they were essentially measuring the same phenomena.

The survey questions were similarly correlated with each other. Special citizen conditions (such as someone present being sick or injured) and special officer actions (such as arrests or the use of force), occurred infrequently. Questions about these conditions or actions were combined, when appropriate, in order to minimize spurious correlations.

---

7 When the officer was patrolling with a partner, the racial characteristics of the police team were categorized as black, white or mixed. The racial characteristics of the team were selected instead of the race of the officer under study, since citizens were more likely to respond to the racial characteristics of both officers than to just one of them.

8 The survey questions which were combined to produce new dummy variables included: (1) the presence of any sick or injured persons when the police arrived (Q. 11 and 12); (2) the use of force, insults, threat of arrest or arrest by police; (3) no dangerous, angry or upset persons present (Q. 14, 17 and 20) [Note: questions 14, 17 and 20 were also included in the analysis as separate questions]; (4) persons present who were dangerous, angry or upset after the police arrived (Q. 15, 18 and 21). (Dummy variables have a value of 1 if the condition listed was present, and a value of zero otherwise.)
Following the first two correlational matrices, one background variable was eliminated because of its high correlation with another, and all of the survey questions (and combinations of questions), which occurred with low frequency were dropped or combined with other questions, since high but statistically non-significant correlations appeared between items occurring with low frequency.

The remaining background variables and survey questions were placed in a third correlational matrix to determine whether the background variables were related to the survey questions. Factor analysis was also performed using the variables in this third matrix. Since the background variables appeared to be unrelated to survey responses, the only group comparisons which were performed were between all women and all men, women patrolling alone vs. men patrolling alone, and women patrolling with partners vs. men patrolling with partners.

COMMUNITY SURVEY

The Community Survey was designed to measure attitudes towards policewomen on patrol. A representative sample of residents in the experimental and comparison districts was interviewed to determine attitudes towards policewomen and expectations of how policewomen would handle various patrol situations.

Instrument Description

The Community Survey was an attitude survey administered by telephone. The survey contained a number of questions with pre-defined categories. Open-ended questions (Why do you say that?), followed several questions about the expected effectiveness of policewomen
compared to policemen in various patrol situations, which were coded into nineteen response categories and three directional categories (positive, neutral and negative comments). There were several measures of the type and degree of contact the citizen had had with a policewoman, including a range from direct contact to exposure from the media. Respondents were asked to estimate the effect of the media and/or of direct contact on their attitudes. Respondents were also asked about the safety of their neighborhood, their general level of satisfaction with police, comparisons of the expected level of respect citizens would show policewomen compared to policemen, the respectfulness of policewomen toward citizens compared to policemen, the expected effect of policewomen on criminal apprehension and the crime rate, and general attitudes toward women on patrol and women's rights.

Sample

The Community Survey was administered three times during 1972 and once in 1973, to assess the effects over time of policewomen on patrol in the E and C districts. The first administration was prior to the time when newly hired women had graduated from training academy, although a few reassigned women were on patrol. Chilton Research Services conducted the telephone interviews and worked in cooperation with The Urban Institute Staff to determine survey specifications. In 1972 both residents and people in the business community were interviewed, but since business persons and residents were found to be very similar in their attitudes, business interviews were discontinued in 1973.
The survey design called for completed interviews from 144 residents during the first administration and 120 residents during each of the following three time periods (waves), with equal numbers of men and women and designated proportions of three age categories in the E and C districts. Forty businesses were surveyed during each of the three administrations in 1972, with roughly equal numbers of high, middle and low crime risk businesses selected for interview (e.g., beauty shops were considered low risk, and liquor stores high risk). Age and race quotas were developed to reflect the population distribution in the survey area. In order to increase the reliability of subsamples, the number of white respondents was increased to roughly twice their proportion in the community sampled (only about 14 percent of the residents in the E and C districts were white), and an attempt was made to obtain identical sample sizes for each age/race/sex group in the E and C areas. Due to the difficulty of filling some of the quotas, the actual cell sizes were slightly at variance with the original design (see Table 14).

Table 14: Community and Business Survey Samples

<table>
<thead>
<tr>
<th>WAVE</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td>E</td>
<td>C</td>
<td>E</td>
<td>C</td>
<td>E</td>
</tr>
<tr>
<td>Community</td>
<td>72</td>
<td>72</td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>Business</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Since all of the Community Survey interviews were conducted by telephone, individuals living in households without telephones were not interviewed. In the first phase of the study (1972), it was difficult to locate enough white respondents in the comparison districts, since only about 9 percent of the comparison residents were white (one of the experimental districts had about 10 percent and the other about 30 percent white residents). Therefore, the following modifications were instituted to increase the likelihood of locating white residents: (1) after all black respondent cells had been filled in the comparison districts, interview attempts were restricted to three telephone exchanges from which at least one interview from a white respondent had been obtained, and (2) after the 170th interview, a sample of households living on streets which had produced at least one white respondent at that point was developed from which interviews were attempted.

Telephone numbers were selected by a computer-generated random number sampling procedure. A screening form was used to determine whether the residence was in the District of Columbia (some exchanges overlapped into Maryland or Virginia) and whether at least one member of the household was 18 or older. The interview was terminated if these two conditions were not met.

Interviewers used six versions of an age/sex grid to insure random selection of the person interviewed (see Figure 2). The six versions were rotated in a systematic fashion to avoid interviewer bias.

In the last administration of the Community Survey (1972), these procedures were modified so that the household street sample was developed after all of the exchanges likely to produce white respondents had narrowed the remaining number of white respondents to be interviewed to 10 or less.
The interviewer determined whether or not there were one or more persons in the household in the selected age/sex group (cell 1). If no persons in the selected group were present, the interviewer selected the next cell (2) on the grid. If there was more than one individual in the selected cell, the youngest person was selected to counteract the tendency of "at home" individuals to be older. Interviewers bypassed cells when particular age and sex quotas were filled. Following interviewee selection, respondents were asked their race. Interviews were terminated if that age/race/sex quota had been filled. Addresses were determined at the close of the interview rather than at the beginning so as to reduce the chance of refusal. Respondents who lived outside the survey area, whose addresses could not be determined or who lived in the districts where the quota had been exceeded were eliminated after the interview was completed.
Analysis

Residents of the E districts were first compared with residents in the C districts. Since there were virtually no district differences, including degree of contact with policewomen, interviews from E and C districts were combined for further analysis. Data was compared by sex, age and race of citizen. The results from the 3 time periods in 1972 were compared, and since they were very similar, the 1972 data was combined. Results from the 1973 period were compared to the 1972 data in order to ascertain whether earlier expectations regarding policewomen had changed over time.

PATROL SURVEY

The Patrol Survey is a self-administered attitude survey which was designed to measure police officers' attitudes toward women on patrol as well as certain job-related attitudes. It was completed by patrol officers of both sexes, including male officers with varying seniority in both the experimental and comparison districts.

Instrument Description

The Patrol Survey is an anonymous self-administered questionnaire which was designed to measure officers' attitudes toward women in policing, toward other patrol officers and officials, toward police work in general, and toward the community. Since the 1972 study showed that men and women differed in the number of moving traffic citations issued and arrests made, questions were added to the Patrol Survey in 1973 to determine whether attitudes towards making arrests and giving moving traffic citations also differed by sex. In order to determine
the reasons for assignment changes, an extra page was added in 1973 (Patrol Survey Insert), in which new women and comparison men were asked to specify their current and past assignments, length of assignment, and, if they were currently in an assignment other than regular patrol duty, to state the reasons for their transfer and their preference for regular patrol versus their current assignment.

Respondents were to indicate their answer by placing a mark on lines in which only the midpoint and the extremes were labeled. Marks on the line were later converted into scaled scores, ranging from 0 on the extreme left to 8 on the extreme right. Whenever possible, "negative" responses were on the left, neutral was the midpoint, and "positive" responses were on the right. This type of scaling was intended to simplify the response possibilities for respondents as well as to simplify data presentation across questions.

The Patrol Survey contained two lists (one for males and one for females), of thirteen relatively common patrol situations, such as "handling disorderly males," and two lists of twelve traits, such as "intelligent." Officers were to indicate the percentage of male and female officers with a year's street experience who could satisfactorily handle each patrol situation and the percentage who they felt possessed each trait. Officers also were asked to check up to five (of the twelve) traits which they considered to be "very important for a police officer."

In order to control for bias due to the order of presentation (i.e., whether the male or female list was presented first), two different forms were used. On Form 1, officers checked the percentages
of males deemed satisfactory on the patrol situations before checking the percentage of females deemed to be satisfactory, and the percentages of females possessing each "trait" was checked before the percentage of males on these "traits." On Form 2, the sex order was reversed on both the "situations" and "traits." In 1972, "form" was found to affect the officers' responses to the "patrol situations" and "traits" questions \(^{10}\) (this analysis was not performed in 1973).

**Sample**

In April, 1972, prior to the time when women were assigned to uniformed patrol, a preliminary version of the Patrol Survey was administered as a pretest to 49 male officers (refer to Bloch, Anderson, & Gervais, 1973a, pp. 1-13 for the pre-test instrument). Thirty-six surveys were returned and useable. In the fall of 1972, a revised version of the Patrol Survey was distributed to 80 (newly hired) policewomen in the E districts, 80 comparison men in the C districts and to 120 "experienced" male officers (having a minimum of one year's experience on the force), 80 of whom were assigned to the E districts and 40 to the C districts, for a total of 280 surveys. All officers surveyed were assigned to the Patrol Division. In August 1973, a

---

\(^{10}\) Officers given Form 1, which had the males listed first for "patrol situations," rated male officers more favorably than did officers given Form 2, where females were rated first. To a lesser degree, respondents given Form 2, where males were listed before females on "traits," tended to say that more males possessed the traits than respondents given Form 1. Ratings of female officers were generally unaffected by "Form." Questionnaires were not analyzed by "form" in 1973, so it could not be determined whether the effect of "form" held up over time. If the effect was similar, the fact that a larger proportion of Form I questionnaires were distributed and returned in 1973 than in 1972 would have biased the findings against the women.
slightly modified version of the Patrol Survey was distributed to all of the new women, to a specially selected group of comparison men assigned to the E and C districts, and to 50 randomly selected men in the C districts and 120 randomly selected men in the E districts. In both years, a larger number of surveys were distributed to randomly selected male officers in the E districts than in the C districts in order to obtain a sizeable sample of experienced officers of various ages who had worked with policewomen.

In 1972, the return rate was 72 percent, compared with 31 percent in 1973 (see Table 15).

Table 15: Patrol Survey Sample

<table>
<thead>
<tr>
<th>Officers</th>
<th>1972 Returned</th>
<th>1972 Distributed</th>
<th>1973 Returned</th>
<th>1973 Distributed</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Women^a</td>
<td>57</td>
<td>80</td>
<td>15</td>
<td>65</td>
</tr>
<tr>
<td>Comparison Men^a</td>
<td></td>
<td>80</td>
<td>23</td>
<td>71</td>
</tr>
<tr>
<td>Randomly Selected Men</td>
<td></td>
<td>120</td>
<td>55</td>
<td>170</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>199</strong></td>
<td><strong>280</strong></td>
<td><strong>98</strong></td>
<td><strong>312</strong></td>
</tr>
</tbody>
</table>

^aIn 1972, no attempt was made to distinguish the comparison men from other men surveyed in the C districts. A total of 57 surveys were returned and useable from C district males in that year. In 1973, new women and comparison men, but not the other officers surveyed, were given "Patrol Survey Inserts," together with the Patrol Survey. Surveys returned with the Inserts were considered to have been completed by the officers under study.
Among the possible reasons for the low response rate in 1973 are: (1) a lack of interest in completing the same survey (a number of officers requested to complete the survey had completed the Patrol Survey previously); (2) negative opinions about the large number of surveys and observations which have been performed in these districts for the policewomen evaluation; (3) doubt about the survey's relevance, since Chief Wilson had already announced the policewomen's program to be a success; (4) lack of interest in expressing opinions about policewomen (the novelty may have diminished); (5) the fact that there was no departmental order regarding completion of this survey in 1973, whereas such an order had been issued in 1972; and (6) lack of payment for completing the survey (officers had been paid $3.00 for completion of the survey in 1972).

Analysis

Male and female officers were compared on all questions, as were black male officers with white male officers. Officials were compared with male and female officers on questions common to the Patrol and Officials' Surveys. Responses from males in 1972 were compared with responses from males in 1973. Males in the experimental districts were compared with males in the comparison districts only on the "difference scores" (male score minus female score), for the patrol situations and traits questions. The sample of female respondents was too small in 1973 to provide reliable data on subgroups within the sample or to permit comparisons over time.
In the first study period, multiple regression analysis was performed on the Patrol Survey in an attempt to account for the variance in the absence of colinear variables. The independent variables included sex, race, the log of total days spent on patrol with a female officer, the percent of total days in the last month spent on patrol with a woman, age, seniority (in months), the log of seniority, district, marital status, children, education, height and "form" of the instrument. Dependent variables included all of the 102 survey questions. Four variants of equations containing these variables were originally run; t-ratio's were calculated for the coefficient of each independent variable, and F-ratios were calculated for the entire equation. Following analysis of these equations, the variables which contributed little to the variance as well as "form" were deleted, and six new equations were run with the remaining variables (see Bloch & Anderson, 1973, pp. 77-79 for a complete description of the various equations used, and ibid, pp. 81-93 for the results of the multiple regression analysis). Also during the first phase of the study, 10 indices were constructed from combinations of questions from this survey (Bloch & Anderson, 1973, pp. 94-95). The results of both the regression analyses and the indices did not warrant further analysis of this type, so this was not continued in 1973.

PERSONNEL JACKETS

Data from the personnel jackets of new women and comparison men were recorded, including background characteristics, selection and training scores, and performance statistics.
Instrument Description

The personnel jackets are the officer's permanent file, which is kept within the department for administrative purposes, rather than a measurement instrument. Selected information from these jackets was recorded, including: 1) demographic and personal history variables (age, height, weight, education, marital status, children, number of previous jobs, and the number of arrests and moving traffic violations prior to appointment); 2) selection and training variables (pre-employment interview scores, police training academy scores, police academy driving test results, and Civil Service [entrance] examination scores); 3) performance statistics (the number of traffic accidents, injuries, and days of light duty since appointment, the number of felony arrests and misdemeanor arrests made since graduation from training academy, instances of unbecoming conduct, and the number of written complaints and commendations from citizens and the department); and 4) the previously mentioned Departmental Ratings.

Sample

In 1972, personnel jackets were obtained for only 58 men and 58 women (out of 80 men and 80 women). Civil Service scores were not available to the evaluators until 1973. In 1973, fairly complete data was eventually obtained for 84 women and 82 men. The analysis for most of this data was performed prior to receiving the last nine jackets. Hence, the actual sample used in these analyses consisted of 79 women and 78 men.
Analysis

New women and comparison men were compared on background characteristics, selection and training scores, and the performance measures indicated above. Data on officers for whom personnel jackets were available in 1973 (the full sample of 84 women and 82 men) were entered into the Super-Personnel File (see below).

SUPER-PERSONNEL FILE

The Super Personnel File was a computer file in which data about individual officers were entered to permit analysis of the relationship between background and performance characteristics.

Sample

Data on 84 new women and 82 comparison men were included in this file, which was the total number of officers for whom personnel jackets could be located. Actual sample sizes for individual variables in the file vary depending upon the source of the information. The following variables were included in this file: 1) all of the personnel jacket data (background characteristics, selection and training scores, Departmental ratings, and all of the performance statistics available in their jackets); 2) all of the Chief's Survey variables (assignments, arrests, days sick leave, light duty, injuries, and officer ratings on 8 patrol skills); 3) average ratings of each officer on each of the 4 patrol skills listed in the Official's Survey ratings; and 4) selected variables from the Structured Observations (incidents per hour, average time per incident, officer-partner leadership patterns, and various officer-citizen interactions by type of citizen).
Analysis

Separate correlational analyses were performed on background and performance variables. When performance variables having similar content were found to be highly correlated with each other, representative variables were selected for inclusion in the later analysis. Variables which were more general, and which would apply to a larger sample of officers, were selected over more specific items. The selected performance variables were then correlated with background variables, within sex. In addition, officers were compared by type of assignment, within sex.

Performance data on 18 officers selected as the "best" and "worst" of their sex on 3 different criteria (Departmental ratings, Official's Survey ratings, and number of misdemeanor arrests) were presented. For each criteria, the average performance data from the 3 men and 3 women who were highest and the 3 men and 3 women who were rated lowest were reported, in order to provide a rough picture of the performance of officers who were thought of as unusually good or poor (for their sex) by the department. This was intended to be purely descriptive, rather than statistical.

ARREST MONITORING

Arrests made by new women and comparison men were examined to determine whether there was a difference in the number, type or outcome of arrests. Since the proportions of different types of arrests were similar, the principal comparisons concerned the disposition of arrests. For each incident in which a felony or misdemeanor arrest
was made, the most serious charge as well as the most serious disposition was recorded. Men and women were compared on their conviction rate as well as other types of arrest outcomes. The reasons for "no-papers" (the least serious disposition, in that the case is rejected outright by the prosecutor, were also analyzed by sex. Refer to Bloch & Anderson, 1974b, for a more detailed description of all procedures utilized in assessing arrest dispositions.

INTERVIEWS

Interviews were conducted with a sample of officers to obtain a more complete picture of the program problems, the differences between male and female officers, and the attitudes of officers of both sexes toward policewomen performing various police functions.

In 1972, interviews were conducted with district commanders in the two E and two C districts about a month before the first newly hired woman graduated from training academy, and at the close of the first year of study, after the women had been under their command for several months. Any new women and comparison men who had been involved in critical incidents were interviewed to determine what had occurred. Reassigned women were interviewed in the police academy, during their week of retraining for patrol duty. At the close of the 1972 data collection period, groups of male and female officers were interviewed in the E and C districts to gather information about their experience with the program. Information about ongoing program developments was obtained more informally from the policewomen's coordinator and the Director of Administrative Services.
In 1973, interviews with male officers in the E and C districts, and female officers in the E districts, were obtained at two points in time. All of these officers were on patrol assignments when selected for interviews. The initial plan called for interviews from 20 officers (9 new women, 4 men in the C districts, and 7 men in the E districts). Only 18 of these officers were available for interviews the first time, and 13 (8 women and 5 men) during the second interview period, due to resignations, injuries, vacations, and refusals. Interview topics included whether policewomen received equal treatment, whether they knew of any incidents which demonstrated policewomen's ability or lack of ability, whether policewomen approached any types of situations differently than policemen, how citizens reacted to policewomen, and during the second interview period, officers who had changed assignments were questioned about the reasons for the change and their reactions to the change.
APPENDIX D

Examples of Measurement Instruments

1. Example of performance observations (Structured Observation)
   a. Profile Sheet
   b. Full Incident form
   c. Brief Incident form
   d. Observer's Ride-Along Summary

2. Example of Service Survey

3. Example of Departmental Rating Form

4. Example of Chief's Survey

5. Example of Official's Survey
   a. Ratings
   b. Attitudes

6. Example of Patrol Survey

7. Example of Community Survey
### STRUCTURED OBSERVATION

#### PROFILE SHEET

<table>
<thead>
<tr>
<th>Observer: Name</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date: Month _ _ Day _ _</td>
<td></td>
</tr>
<tr>
<td>District _ _ Beat _ _ Day 1. Mon - Thurs</td>
<td></td>
</tr>
<tr>
<td>Officer #1: Name</td>
<td>Code</td>
</tr>
<tr>
<td>Officer #2: Name</td>
<td></td>
</tr>
</tbody>
</table>

| 1. Sex: 1. Male 2. Female |
| 2. Age: (years) |
| 4. Height: (inches) |
| 6. Months on MPDC |
| 7. Last Grade Completed |

| 12. High school or GED |
| 13. 1 year additional school |
| 14. 2 years additional school |
| 15. 3 years additional school |
| 16. College degree (B.A., B.S.) |
| 17. 1 year graduate school |
| 18. 2 years graduate school (M.A.) |
| 19. 3 years graduate school (M.A.) |
| 20. 4 years graduate school (Ph.D., M.D.) |

| 8. Is either member permanently assigned to this beat? |
| 1. Officer #1 2. Officer #2 3. Both Officers |

**TO BE ASKED OF OFFICER #1 ONLY (if applicable)**

| 9. How many times have you worked with a policewoman on uniformed patrol before? (Do not count undercover, other assignments) |

(code more than 99 times as 99 in box at right)

<table>
<thead>
<tr>
<th>CARD #1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>14</td>
</tr>
</tbody>
</table>
(Performance Observations, cont.)

Full Incident Form

<table>
<thead>
<tr>
<th>ADVICE/ASSIST (continued)</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
</tr>
</thead>
<tbody>
<tr>
<td>j. Served warrant</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Officer #1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Officer #2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k. Assisted (e.g., recover possessions)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Officer #1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Officer #2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>l. Investigated - no report</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Officer #1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Officer #2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m. Transported - voluntary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Officer #1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Officer #2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n. Took report</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Officer #1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Officer #2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. ACTION

| o. Frisked or searched person |    |    |    |    |    |
|    Officer #1                |    |    |    |    |    |
|    Officer #2                |    |    |    |    |    |
| P. Transported - involuntary|    |    |    |    |    |
|    Officer #1                |    |    |    |    |    |
|    Officer #2                |    |    |    |    |    |
| q. Searched premises/vehicle (investigative search) |    |    |    |    |    |
|    Officer #1                |    |    |    |    |    |
|    Officer #2                |    |    |    |    |    |
| r. Threatened arrest        |    |    |    |    |    |
|    Officer #1                |    |    |    |    |    |
|    Officer #2                |    |    |    |    |    |
| s. Threatened physical force|    |    |    |    |    |
|    Officer #1                |    |    |    |    |    |
|    Officer #2                |    |    |    |    |    |
| t. Threatened use of weapon |    |    |    |    |    |
|    Officer #1                |    |    |    |    |    |
|    Officer #2                |    |    |    |    |    |
| u. Arrested                 |    |    |    |    |    |
|    Officer #1                |    |    |    |    |    |
|    Officer #2                |    |    |    |    |    |
BRIEF INCIDENT SHEET

(USE ONLY IF THERE WAS NO CITIZEN CONTACT, OR BRIEF CONTACT, BUT NO SIGNIFICANT ACTIONS TAKEN BY OFFICERS OR CITIZENS)

Observer: Name__________________________ Code__________________________

Date: Month__________Day__________ Incident #__________________________

A. Number of minutes between start of incident and time back in service__________

1c. IF A TRAFFIC INCIDENT, (do not record parking tickets) describe actions of officers. If more than one action occurred, code priorities are: 4, 1, 2, 3, 5, 7, (e.g., if 4 & 3 occurred, you would only code 4)

Officer #1 Officer #2

1. gave traffic ticket ........................................
2. gave warning, no ticket ....................................
3. spot check for stolen auto(motorcycles, bikes, etc.) .............
4. arrested person ...........................................
5. other (describe) _________________________________________
6. did not observe enough to judge . . . ._____________________
7. checked license or registration . . . ._____________________

1d. IF AN INCIDENT OTHER THAN TRAFFIC WAS INITIATED BY OFFICERS, enter number of incident________

1. talk with juveniles
2. talk with adults
3. question suspicious person(s)
4. follow-up of prior incident (describe)
5. business or bank check (talk to businessmen)
6. business or school check for break-in (unoccupied)
7. check of known criminal areas
8. Other incidents (community relations or preventive patrol) not recorded in Q.2. (describe).

3. IF POLICE LEFT SCENE without performing any activity, why did they leave? (enter code)____________

1. wrong address
2. unable to locate any incident (nothing found)
3. police were not wanted
4. other police handled situation
5. Other (specify)________________________
(Performance Observations, cont.)

Observer's Summary
(completed after last observation tour)

Ride-Along Summary

6. During this study, we discovered that a large number of women were assigned to the station, wagon or foot beats. Do you have any knowledge of why this was the case? (E.g., did any women you spoke to request this, prefer it, dislike station, wagon, etc., and what were the circumstances of their being assigned to these other duties?) Specify duty assigned to.

7. How did the citizens react to policewomen? (Distinguish male and female citizens - this questions may include citizens on the street as well as in the incidents.)

8. a. Did you see any incidents which a woman handled particularly well or very poorly? Please describe.
   b. Did officers or officials mention any critical incidents which had been important in forming their opinions of policewomen? If so, describe.
### SERVICE SURVEY
(Citizen Interviews)

13. (ASK ONLY IF YES TO Q. 11 OR Q. 12) If you had to rate that particular officer, how good a job did he/she do in helping the sick or injured people? Would you say he/she did as:

<table>
<thead>
<tr>
<th>Rating</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very good job</td>
<td>1</td>
</tr>
<tr>
<td>Good job</td>
<td>2</td>
</tr>
<tr>
<td>Average job</td>
<td>3</td>
</tr>
<tr>
<td>Poor job or</td>
<td>4</td>
</tr>
<tr>
<td>Very poor job</td>
<td>5</td>
</tr>
<tr>
<td>DON'T READ</td>
<td></td>
</tr>
</tbody>
</table>

14. We are interested in how police deal with dangerous, angry or upset people. Was there anyone there who was dangerous?

<table>
<thead>
<tr>
<th>Response</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>No</td>
<td>2</td>
</tr>
<tr>
<td>Don't Know</td>
<td>8</td>
</tr>
</tbody>
</table>

15. Did they become dangerous before the police got there or sometime after the police got there?

<table>
<thead>
<tr>
<th>Time</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>1</td>
</tr>
<tr>
<td>After</td>
<td>2</td>
</tr>
<tr>
<td>Don't Know</td>
<td>8</td>
</tr>
</tbody>
</table>

16. Why did they become dangerous? (PROBE)

17. Was there anyone there who was angry but not dangerous?

<table>
<thead>
<tr>
<th>Response</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>No</td>
<td>2</td>
</tr>
<tr>
<td>Don't Know</td>
<td>8</td>
</tr>
</tbody>
</table>

18. Did they become angry before the police got there or sometime after the police got there?

<table>
<thead>
<tr>
<th>Time</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>1</td>
</tr>
<tr>
<td>After</td>
<td>2</td>
</tr>
<tr>
<td>Don't Know</td>
<td>8</td>
</tr>
<tr>
<td>Rating</td>
<td>Unsatisfactory</td>
</tr>
<tr>
<td>--------</td>
<td>----------------</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

1. Bearing and behavior
2. Human Relations
3. Learning ability
4. Knowledge and skill
5. Acceptance of Responsibility
6. Written expression
7. Oral expression
8. Performance of duty
9. Overall evaluation

10. Do you have any knowledge of unbecoming conduct:
   1) Yes ______  0) no ______

IF YES, Explain: ____________________________________________
                                                                 ____________________________________________
                                                                 ____________________________________________
                                                                 ____________________________________________
<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Arrests for Felonies:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Arrests for Misdemeanors:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Moving Traffic Violations:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Favorable Letters or Comments from the Public:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Unfavorable Letters or Comments from the Public:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Certification to Drive:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Certification to Patrol Alone:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IF NOT, ESTIMATED DATE OF CERTIFICATION:</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**SAMPLE QUESTION:**

<table>
<thead>
<tr>
<th>Question</th>
<th>Extremely poor considering length of street experience</th>
<th>Average</th>
<th>Extremely good considering length of street experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Dealing with the Public</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Handling a Disorderly Male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Handling a Public Fight</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Card #2

260
OFFICIAL'S SURVEY
(Includes example of ratings)

106. How would you improve the policewomen's program? PLEASE EXPLAIN YOUR ANSWER ON THE BLANK PAGES AT THE BACK OF THIS BOOKLET.

---------------------------------------------------------------------------------------

FOR SERGEANTS ONLY

107. During the last month, approximately how many officers in your section were men? _______________

108. During the last month, approximately how many officers in your section were women? _______________

ENCLOSED IS A LIST OF PROBATIONARY OFFICERS WHO WORK IN YOUR DISTRICT. PLEASE RATE ONLY THOSE OFFICERS THAT WORK IN YOUR SECTION AND THAT YOU KNOW WELL ENOUGH TO JUDGE THEIR COMPETENCE IN THE FOUR PATROL SITUATIONS LISTED.

---------------------------------------------------------------------------------------

FOR LIEUTENANTS AND CAPTAINS ONLY

107. During the last month, approximately how many officers under your command were men? _______________

108. During the last month, approximately how many officers under your command were women? _______________

ENCLOSED IS A LIST OF PROBATIONARY OFFICERS WHO WORK IN YOUR DISTRICT. PLEASE RATE ONLY THOSE OFFICERS THAT WORK UNDER YOUR COMMAND AND THAT YOU KNOW WELL ENOUGH TO JUDGE THEIR COMPETENCE IN THE FOUR PATROL SITUATIONS LISTED.

---------------------------------------------------------------------------------------

FOR ALL OFFICIALS

For those officers you have just been asked to rate, please consider their length of street experience in your rating. Indicate your ratings by drawing a vertical line in the appropriate place (see example below). For example, if you are familiar with Officer Jones' job performance, you would skip to "Officer Jones" on the enclosed list, then rate how Officer Jones generally performs in each of the four patrol situations. DO NOT RATE ANY OFFICERS UNLESS YOU ARE FAMILIAR WITH THEIR PERFORMANCE.

EXAMPLE

<table>
<thead>
<tr>
<th>Handling Domestic fights</th>
<th>Handling street situations involving violence or potential violence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Officer Jones</td>
<td></td>
</tr>
<tr>
<td>unsatisfactory average excellent unsatisfactory average excellent</td>
<td></td>
</tr>
</tbody>
</table>
A patrol officer does many different kinds of things while on duty. Below are a list of things an officer might do. In your opinion, what percent of female patrol officers with a year's experience would be able to handle the following situations satisfactorily? Please indicate by placing a mark on the line at the appropriate place.

73. Handling the victims of an armed robbery
   □ 0%  □ 50%  □ 100%

74. Dispersing a group of noisy juveniles (ages 9-12)
   □ 0%  □ 50%  □ 100%

75. Handling a down and out drunk
   □ 0%  □ 50%  □ 100%

76. Handling disorderly males
   □ 0%  □ 50%  □ 100%

77. Arresting prostitutes
   □ 0%  □ 50%  □ 100%

78. Handling traffic accidents
   □ 0%  □ 50%  □ 100%

79. Questioning a rape victim
   □ 0%  □ 50%  □ 100%

80. Cruising around and observing
   □ 0%  □ 50%  □ 100%

81. Handling threatening situations where someone has a knife or a gun
   □ 0%  □ 50%  □ 100%

82. Handling disorderly females
   □ 0%  □ 50%  □ 100%

83. Writing Reports
   □ 0%  □ 50%  □ 100%
**PATROL SURVEY**

**WHAT IS YOUR REACTION TO EACH OF THE FOLLOWING THREE STATEMENTS?**

26. A good leader should be strict with people under him in order to improve their performance.

<table>
<thead>
<tr>
<th>Strongly disagree</th>
<th>Neutral</th>
<th>Strongly agree</th>
</tr>
</thead>
</table>

27. Respect for the police in a tough neighborhood depends on the willingness to use force frequently and effectively.

<table>
<thead>
<tr>
<th>Strongly disagree</th>
<th>Neither agree</th>
<th>Strongly agree</th>
</tr>
</thead>
</table>

28. Some force is necessary and justified when a citizen unjustly insults and curses a police officer.

<table>
<thead>
<tr>
<th>Strongly disagree</th>
<th>Neither agree</th>
<th>Strongly agree</th>
</tr>
</thead>
</table>

29. About what percent of the Department's male patrol officers are highly motivated in their job and do even more than is required? __% 

30. Is there one sergeant assigned to your section that you regularly talk to about your job and your job problems? Yes ___ No ___

31. How much do your sergeants know about how well you do your job? 

<table>
<thead>
<tr>
<th>Nothing</th>
<th>About half</th>
<th>Everything</th>
</tr>
</thead>
</table>

32. If you should have complaints about your job, about what percent of the time would you expect your officials to be understanding? 

<table>
<thead>
<tr>
<th>0%</th>
<th>50%</th>
<th>100%</th>
</tr>
</thead>
</table>

33. How do you feel about doing scout car patrol?

<table>
<thead>
<tr>
<th>Dislike it</th>
<th>Neutral</th>
<th>Like it</th>
</tr>
</thead>
<tbody>
<tr>
<td>very much</td>
<td></td>
<td>very much</td>
</tr>
</tbody>
</table>

34. How do you feel about patrolling on a foot beat?

<table>
<thead>
<tr>
<th>Dislike it</th>
<th>Neutral</th>
<th>Like it</th>
</tr>
</thead>
<tbody>
<tr>
<td>very much</td>
<td></td>
<td>very much</td>
</tr>
</tbody>
</table>

35. About what percent of the Department's female patrol officers are highly motivated in their job and do even more than is required? __%
17. Suppose a fight broke out among several people on a street corner and half of the police officers sent were policewomen rather than all being policemen. Do you think the fact that half of the officers were policewomen would make the situation ... much more likely to get out of hand, somewhat more likely, no difference, somewhat less likely, or much less likely to get out of hand?

<table>
<thead>
<tr>
<th>Response</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much more likely</td>
<td>1</td>
</tr>
<tr>
<td>Somewhat more likely</td>
<td>2</td>
</tr>
<tr>
<td>No difference</td>
<td>3</td>
</tr>
<tr>
<td>Somewhat less likely</td>
<td>4</td>
</tr>
<tr>
<td>Much less likely</td>
<td>5</td>
</tr>
</tbody>
</table>

18. Suppose a riot broke out somewhere in the District and half of the police officers sent were policewomen rather than all being policemen. Do you think the fact that half of the officers sent were policewomen would make the situation ... much more likely to get out of hand, somewhat more likely, no difference, somewhat less likely, or much less likely to get out of hand?

<table>
<thead>
<tr>
<th>Response</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much more likely</td>
<td>1</td>
</tr>
<tr>
<td>Somewhat more likely</td>
<td>2</td>
</tr>
<tr>
<td>No difference</td>
<td>3</td>
</tr>
<tr>
<td>Somewhat less likely</td>
<td>4</td>
</tr>
<tr>
<td>Much less likely</td>
<td>5</td>
</tr>
</tbody>
</table>

19. How good a job would a policewoman do in investigating a crime scene in order to find evidence — compared to a policeman? Do you think a policewoman would do ... much worse, somewhat worse, about the same, somewhat better, or much better than a policeman?

<table>
<thead>
<tr>
<th>Response</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much worse</td>
<td>1</td>
</tr>
<tr>
<td>Somewhat worse</td>
<td>2</td>
</tr>
<tr>
<td>About the same</td>
<td>3</td>
</tr>
<tr>
<td>Somewhat better</td>
<td>4</td>
</tr>
<tr>
<td>Much better</td>
<td>5</td>
</tr>
</tbody>
</table>

20. Now, think about the number of crimes committed in the District of Columbia last year. If half of the police officers had been policewomen, do you think the number of crimes would have been ... much higher, somewhat higher, about the same, somewhat lower, or much lower than the actual number of crimes committed with all policemen?

<table>
<thead>
<tr>
<th>Response</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much higher</td>
<td>1</td>
</tr>
<tr>
<td>Somewhat higher</td>
<td>2</td>
</tr>
<tr>
<td>About the same</td>
<td>3</td>
</tr>
<tr>
<td>Somewhat lower</td>
<td>4</td>
</tr>
<tr>
<td>Much lower</td>
<td>5</td>
</tr>
</tbody>
</table>

DO NOT READ: Don't Know
BIBLIOGRAPHY


Armat, V. Policewomen on patrol. The Reader's Digest, July 1975. (Condensed from The Saturday Evening Post).


Civil Rights Act of 1964, Title VII (78 Statute 253; 42 U.S. Congress, Section 2000c-12.


Roos, N. P. Contrasting social experimentation with retrospective evaluation: a health care perspective. Public Policy, Spring 1975, 23 (2), 241-258.


