T.D. 2/96
Decision rendered on February 15, 1996

THE CANADIAN HUMAN RIGHTS ACT
(R.S.C., 1985, c. H-6 (as amended))

HUMAN RIGHTS TRIBUNAL

BETWEEN:

PUBLIC SERVICE ALLIANCE OF CANADA

Complainant

- and -

CANADIAN HUMAN RIGHTS COMMISSION

Commission

- and -

TREASURY BOARD

Respondent

DECISION OF THE TRIBUNAL

Tribunal: Donna Gillis, Chairperson
Norman Fetterly, Member
Joanne Cowan-McGuigan, Member

Appearances: Andrew Raven
Counsel for the Public Service Alliance of Canada

Rosemary Morgan and René Duval
Counsel for the Canadian Human Rights Commission

Duff Friesen, Lubomyr Chabursky and Deborah Smith
Counsel for Treasury Board

Location of Hearing: Ottawa, Ontario

TABLE OF CONTENTS

I. INTRODUCTION

1. The Canadian Human Rights Commission (the "Commission") is established under the Canadian Human Rights Act, R.S., 1985, c. H-6 as amended (the "Act"), and is a party in this complaint, representing the public interest.

2. The Commission presented six witnesses qualified to testify as experts. The first witness to appear was Dr. Nan Weiner, an expert in pay equity and compensation. The second expert to testify was Norman D. Willis, an expert in pay equity and job evaluation. They were followed by two expert statisticians, Dr. Richard Shillington, an expert in data analysis and Alan Sunter, an expert in statistics. Also called were two employees of the Commission, Paul Durber and James Sadler. Durber is an expert in pay equity, job evaluation and other general areas of job evaluation and Sadler is an expert in pay equity and job evaluation.

3. The Respondent, Treasury Board (the "Employer"), is the employer of employees who work in Federal Public Service of Canada listed in

Schedule 1, Part 1 of the Public Service Staff Relations Act, 1966-67, c. 72, s.1, p. 35, Schedule 1 (the "PSSRA"). In addition to Willis, the Employer only called one expert to testify, Fred Owen. Owen was a former Willis consultant and an expert in pay equity and job evaluation.

4. The Complainant, the Public Service Alliance of Canada (the "Alliance"), is an "employee organization" within the meaning of the PSSRA. The Alliance has been certified by the PSSRA to act as bargaining agent for a number of bargaining units in the Federal Public Service. The Alliance is the third largest union in Canada representing approximately 170,000 employees, 70 per cent of whom work outside of the National Capital Region. The Alliance is composed of 18 components which are, with the exception of one or two components, male-dominated. The largest bargaining unit represented by the Alliance is the Clerical and Regulatory Group (the "CR Group") which consists of approximately 50,000 employees. This bargaining unit is 80 per cent female and includes employees performing an extremely wide range of functions.

5. The Alliance called four experts to testify during the course of this hearing. The first was Dr. Pat Armstrong, accepted by the Tribunal as an expert in job evaluation and pay equity. The Alliance also called Dr. Eugene Swimmer, an expert in labour economics and statistics. The Tribunal accepted one Alliance employee, Margaret Jaekl, as an expert in pay equity and job evaluation. Another individual, Margaret I. Krachun, who at the time of the hearing was employed by the Alliance, was accepted as a layperson with some experience in evaluation gained while a member of one of the evaluation committees.

6. The case originally before the Tribunal arose from complaints filed by both the Alliance and the Professional Institute of the Public Service of Canada (the "Institute") alleging violation of s. 11 of the Act. The Institute called one expert witness, Dan Butler, a negotiator with the Institute. He was accepted by the Tribunal as an expert expressing the

opinion of the Institute on several issues before the Tribunal, primarily on wage adjustment methodology.

7. The human rights complaints before the Tribunal now pertain only to the complaints of the Alliance. The Institute's complaints are no longer before us. Those complaints were resolved by a negotiated settlement between the Employer and the Institute. A Consent Order was issued by the Tribunal dated May 31, 1995, giving effect to their settlement.

8. In the case of the Alliance, two complaints remain for our determination. The first complaint, dated December 19, 1984, alleges discriminatory practice contrary to ss. 7, 10 and 11 of the Act with respect to employees in the female-dominated CR Group. It is only the s. 11 portion of the 1984 CR Group complaint which has been referred to the Tribunal for ruling. The complaint presented on behalf of the employees in the CR Group affects the rights of approximately 50,000 workers who belong to this group.

9. The second complaint, dated February 16, 1990, alleges the results obtained through the process of the Joint Union-Management Initiative on Equal Pay for Work of Equal Value has demonstrated the existence of wage rates which are in contravention of s. 11 of the Act with respect to employees in the female-dominated occupational groups: Clerical and Regulatory; Secretarial, Stenographic and Typing; Data-Processing; Educational Support; Hospital Services; and Library Science. This complaint of the Alliance was filed with the Commission shortly after the breakdown of the Joint Union-Management Initiative (which will be detailed later). That complaint relies upon the job evaluation data generated by a study resulting from this initiative claiming, in support of its position, that employees in the identified complainant groups continue to suffer wage rate discrimination contrary to s. 11 of the Act, notwithstanding unilateral payments announced by the Employer in January of 1990.

10. From the outset, the Alliance's preferred position was to attempt to resolve equal pay issues through negotiations with the Employer at the bargaining table. It was only when these measures failed to lead to corrective action that the complaint mechanism of the Act was invoked.

11. The human rights complaints of the Alliance are not the first s. 11 complaints the Alliance has presented under the Act. The earlier complaints include the complaint of the Library Science Group (the "LS Group") and the Hospital Services Group (the "HS Group") on behalf of employees in the female-dominated sub-groups in the General Services Group (the "GS Group").

12. In each of these cases, monetary compensation in the form of wage adjustments were paid to affected employees. The LS Group complaint was resolved with the understanding that final corrective action would await the outcome of the study. In the matter of the HS Group complaint, which was the subject of a Tribunal Order of July 15, 1987, another earlier tribunal, it was expressly understood by the parties that s. 11 complaint

would likewise await final wage gap computations after the conclusion of the study.

13. Each Federal Public Service employee occupies a position which is classified in accordance with the Employer's classification system. The Employer's classification system is comprised of 69 occupational groups, each with its own classification standard ("job evaluation system").

14. In the classification system, positions are classified as belonging to occupational groups, sub-groups (where applicable) and levels. Occupational groups are designated by two-letter abbreviations; sub-groups by three-letter abbreviations. A position is the smallest organizational unit and represents a unique set of tasks and duties performed by an individual. The Employer has the same number of positions as it has employees. On the other hand, a job in the Federal Public Service is a grouping of positions which have the same key duties and responsibilities.

15. The occupational groups are assembled into six occupational categories as follows: (i) the Scientific and Professional Category; (ii) the Administrative and Foreign Service Category; (iii) the Technical Category; (iv) the Administrative Support Category; (v) the Operational Category; and (vi) the Executive Category.

16. In March of 1985, the government initiated pro-active measures to implement the principles of equal pay for work of equal value in the Federal Public Service. It invited unions and management to participate as partners in a senior level Joint Union-Management Initiative (the "JUMI"). The JUMI was directed by a committee (the "JUMI Committee"). The JUMI Committee was asked to prepare a detailed implementation plan in the area of equal pay for work of equal value. The unions, not only the Alliance, but other unions as well, accepted the government's invitation. The Alliance, at the time of accepting this invitation, had established a consistent policy of supporting the principle of equal pay for work of equal value. At the time of the voluntary initiative, there were three outstanding complaints before the Commission under s.11 of the Act.

17. The action plan agreed to by the JUMI Committee was to conduct a study (the "JUMI Study") pursuant to s. 11 of the Act to determine the degree of sex discrimination in pay and to devise methods for system wide correction in order to eliminate sexually based wage disparities (Exhibit HR-11A, Tab 9, Annex B). The Commission was invited to be a participant of the JUMI Study to fulfil the role of an observer at committee meetings and to provide interpretation and guidance when required by the JUMI Committee. (Exhibit HR-11A, Tab 7). The Commission held all s. 11 complaints, which had been filed before the JUMI Study commenced, in abeyance. The

Commission agreed that any new complaints received during the JUMI Study, which might be affected by the study, were to be held in abeyance as well.

18. The JUMI Committee had equal representation from the Employer and eight different unions. The JUMI Committee's first task was to define the parameters of the JUMI Study. Pivotal to its operation was the requirement for joint agreement between management and union representatives on the process to be used during the JUMI Study (the "JUMI Process"). Neither the

unions nor management was to act independently or make decisions in the course of the JUMI Study without joint approval. The JUMI Committee hired Willis & Associates, a consulting firm based in Seattle, Washington, to assist in the Study. Willis & Associates was founded and directed by Norman Willis.

19. Early on in the JUMI Study, the JUMI Committee made it abundantly clear to Willis that he had no decision-making authority in the conduct the JUMI Study. Willis' role was to attend the meetings and to give advice at the request of the JUMI Committee.

20. The JUMI Committee established sub-committees at various stages which were called upon by the JUMI Committee to provide advice, to perform certain tasks, and make recommendations to the JUMI Committee with respect to particular issues. Agreement by members of the JUMI Committee was required in order to form a sub-committee. Each sub-committee thus formed had equal representation from union and management sides.

21. In the fall of 1987, the JUMI Committee established the Equal Pay Study Secretariat (the "EPSS") to conduct the administrative work associated with the JUMI Study. The EPSS was managed by a Treasury Board representative, Pierre Collard. The objective of the EPSS was to provide administrative support to the multiple evaluation committees in the JUMI Study and it was responsible for the coordination of all support activities.

22. In addition to hiring Willis & Associates, the JUMI Committee eventually agreed on other important matters. The JUMI Committee agreed to evaluate positions from male- and female-dominated occupational groups using a common evaluation plan. A comparison of wages paid to male- and female-dominated occupational groups performing work of equal value could then be made. The JUMI Committee agreed the study would be "position specific" using a representative sample of positions. A position-specific study means every different job selected for evaluation is evaluated

separately as opposed to "predominant use" studies in which positions are selected for evaluations that best represent a classification or grouping of jobs. The JUMI Committee agreed only positions from male- and female-dominated occupational groups, as defined in s. 13 of the Equal Wages Guidelines (the "Guidelines"), were to be included in the representative sample.

23. As of March, 1985, based on s. 13 of the Guidelines (which prescribes the criteria defining sex predominance), the parties agreed there were 9 female-dominated occupational groups, 53 male-dominated occupational groups and 8 gender-neutral occupational groups. For clarity, s. 13 of the Guidelines is reproduced as follows:

> 13. For the purpose of section 12, an occupational group is composed predominantly of one sex where the number of members of that sex constituted, for the year immediately preceding the day on which the complaint is filed, at least
>
> 5
>
> (a) 70 per cent of the occupational group, if the group has less than 100 members;
> (b) 60 per cent of the occupational group, if the group has from 100 to 500 members; and
> (c) 55 per cent of the occupational group, if the group has more than 500 members.

24. The nine female-dominated occupational groups represented by the Alliance and the Institute with their abbreviations are listed below:

- Clerical and Regulatory (CR);
- Data Processing (DA);
- Education Support (EU);
- Home Economics (HE);
- Hospital Services (HS);
- Library Science (LS);
- Nursing (NU);
- Occupational and Physical Therapy (OP); and
- Secretarial, Stenographic, Typing (ST).

25. Positions from gender-neutral occupational groups or the Executive Category were excluded from the study. The proposed JUMI Study, although service-wide in nature, was not intended to cover all employees providing services for the Government of Canada. The JUMI Study did not include employees of Crown Corporations nor did it include employees of

separate employers. For purposes of the legislation, separate employers are identified in Part II of the PSSRA as follows:

- Atomic Energy Control Board
- Canadian Advisory Council on the Status of Women
- Canadian Security Intelligence Service
- Communications Security Establishment, Department of National Defence
- Economic Council of Canada
- Medical Research Council
- National Film Board
- National Research Council of Canada
- Natural Sciences and Engineering Research Council
- Northern Canada Power Commission
- Northern Pipeline Agency
- Office of the Auditor General of Canada
- Public Service Staff Relations Board
- Science Counsel of Canada
- Social Sciences and Humanities Research Council
- Staff of the Non-Public Funds, Canadian Forces

26. The sample eventually drawn was representative of positions by groups and levels for female-dominated occupational groups and by group for male-dominated occupational groups. Approximately 2,800 positions from female-dominated occupational groups and 1,500 positions from male-dominated occupational groups were ultimately included in the sample. The sample size and composition met with the approval of Statistics Canada.

27. The JUMI Committee agreed to use the Willis Job Evaluation Plan, with some amendments, as the appropriate job evaluation instrument for evaluating the representative sample of positions. The JUMI Committee also agreed to use the Willis Questionnaire, with amendments, to gather information on the positions to be evaluated. A communications strategy was recommended and agreed upon by a JUMI sub-committee to encourage selected incumbents to participate in the JUMI Study and to provide information on their positions. Position information was then collected from September, 1987 until January 1989.

28. The JUMI Committee acting on Willis' advice established, as a first step in the process of evaluation, a Master Evaluation Committee (the "MEC"). The MEC was asked to evaluate 503 position questionnaires which were to serve as benchmarks and as a frame of reference for all subsequent evaluations by other evaluation committees. The MEC began its important

task in September, 1987 and finished it in July, 1988. In the final analysis, the MEC completed 501 benchmark evaluations.

29. After the MEC completed their evaluations, the remaining evaluations were done by 14 evaluation committees, (the "multiple evaluation committees"). The first five multiple evaluation committees began evaluating in September, 1988. By April, 1989, they had evaluated approximately 1,283 positions. In April, 1989, the multiple evaluation committees were expanded from five to nine. The nine committees included some members of the first five multiple evaluation committees as well as new members. The expanded committees evaluated approximately 1,400 positions between April, 1989 and September, 1989.

30. In May of 1989, the JUMI Committee decided, in view of the slow progress at which the questionnaires were being evaluated, that the sample size should be reduced by approximately 880 positions. The JUMI Committee then agreed to reduce the original sample from 4,300 positions to approximately 3,280 positions. The Office of the Chief Statistician for Statistics Canada was advised of the nature and reasons for the reduction in the sample size and approved the reduction. In the end, the MEC and the 14 multiple evaluation committees evaluated 3,185 positions from the reduced sample of positions.

31. The Commission's representatives functioned as observers throughout the evaluations of the MEC and the multiple evaluation committees. They were present during the meetings of the JUMI Committee and meetings of the multiple evaluation committee chairpersons.

32. Overall, the JUMI process had a number of shortcomings, largely due to the manner in which it operated. According to Willis, the JUMI Committee was "ill-formed". Rather than working as a team, the JUMI Committee functioned in a negotiating mode with the unions on one side and the Employer on the other. As described by Willis, each side spoke with one voice. Because the Employer represented a singular position, this required the unions to caucus in order to respond in one voice. Rather than a joint union-management committee working together as a team, the proceedings were akin to union-management bargaining.

33. As a result, many decisions took a great deal of effort and time and were not easily or amicably achieved. For example, after the first JUMI Committee meeting which was held on September 16, 1985, it took until September 22, 1986, one year later, for the parties to reach an agreement on the Terms of Reference and Action Plan for the JUMI Study.

34. The length of time needed to carry out the JUMI Study prompted the Chief Commissioner of the Commission, on different occasions, to urge the President of the Treasury Board to resolve the outstanding issues occupying the JUMI Committee.

35. Another problem in the JUMI process was the inability of the management and union sides to reach closure on some major aspects of the JUMI Study. For example, when the MEC had completed its benchmark evaluations, Treasury Board withheld whole-hearted support of those evaluations. Although Treasury Board agreed to proceed with the rest of the evaluations, it continued to harbour doubts and indicated its intention to study the reliability of the MEC benchmarks independently.

36. Problems also arose during the course of the multiple evaluation committees' evaluations. Willis recommended disbanding one of the original five multiple evaluation committees. The JUMI Committee rejected this recommendation and could not agree on a resolution. In addition, there were some multiple evaluation committee challenges to the MEC benchmark evaluations. The JUMI Committee established a smaller version of the MEC (the "Mini-MEC") to review and discuss these challenges. The Mini-MEC could not reach a consensus so in the end the matter was never fully resolved.

37. The JUMI Study was intended to encompass four phases. These phases were to be as follows:

Phase I

Agreement on the common evaluation plan to be used to determine the relative value of jobs and on the evaluation of benchmark positions.

Phase II

Agreement on the statistical methodology for sampling actual positions.

Phase III

Sampling and evaluation of actual positions, using the agreed to evaluation plan with benchmarks.

Phase IV

Determination of the degree of wage disparity and recommendations on corrective measures. These may include recommendations to resolve discriminatory aspects of the classification system which contribute to wage inequity as defined in Section 11 of the Canadian Human Rights Act.

(Exhibit HR-11A, Tab 9)

38. During the life of the JUMI Study tension between the management and union sides persisted and intensified. There was disagreement between the union and management sides relating to the release of evaluation scores. The JUMI Committee agreed the data would be released after two-thirds of the evaluations were completed. According to Willis, following the release of the MEC evaluation scores on July 13, 1988, relationships in the JUMI Committee began to deteriorate. It then became apparent to Willis that the climate of the JUMI Committee had changed. When the MEC results were made available to the parties, the Employer's classification system became an issue for the Employer. Willis was troubled and mystified by correspondence he received from the management co-chair on August 18, 1988, which indicated the parties were not ad idem on the purpose of the JUMI Study.

39. During the last few months of the JUMI Study, an issue arose between the union and management sides relating to a report released by Willis & Associates concerning re-evaluations by a Willis consultant of 222 multiple committee evaluations. This issue was never resolved by the JUMI Committee and eventually, led to the final breakdown of the JUMI Study.

40. The parties had contemplated eventual agreement upon a joint recommendation to the President of Treasury Board for implementation of pay equity. Phase 4 of the JUMI Study was never achieved. After approximately four years, in December, 1989, the union side withdrew from the JUMI Study on a temporary basis. In January, 1990, the largest participant union in the JUMI Study, the Alliance, permanently withdrew from the JUMI Study.

41. Early in 1990, the Government of Canada made a decision to unilaterally implement immediate measures to achieve equal pay for work of equal value for female-dominated occupational groups in the Federal Public Service. The measures adopted by the government were based on the evaluation results of the JUMI Study with corrective adjustments for gender bias arising from the controversial report by Willis & Associates on the 222 re-evaluations. Those measures were referred to as the public service "equal pay adjustments" or the "equalization payments". The equalization payments were applied to three female-dominated occupational groups, the CR, NU, and ST Groups.

42. Neither the Commission nor the Alliance, or any of the other participant unions, were consulted by the Employer prior to making these voluntary adjustments. The parties were first informed of the Employer's decision when the President of the Treasury Board made an announcement on January 26, 1990. The adjustments involved payments of approximately $317

million for wages retroactive to April 1, 1985 and payments of $76 million annually in continuing adjustments. The lump sum payments by the government were made retroactive to March 31, 1985, the month in which the Treasury Board President first announced the establishment of the Joint Union-Management Committee to study how gender based wage discrimination would be eliminated in the Federal Public Service.

43. After the breakdown of the JUMI Study, the Commission and the Alliance made it clear to the Employer the data generated by the JUMI Study would be presented as evidence to a Human Rights Tribunal.

44. The formal investigation of the s. 11 complaints lodged with the Commission commenced following the announcement of the equalization payments. Included in its investigation, was an examination by the Commission of the equalization payments. This exercise was done to ensure full adherence to the Act and Equal Wages Guidelines. Following a formal six month investigation, the Commission decided to refer the s. 11 complaints to a Tribunal. That decision was made on October 16, 1990.

45. During the course of this hearing, when the Commission attempted to introduce the JUMI data into evidence, it was met with the objection by the Employer that the data was inadmissible on the grounds it had been created in an effort to resolve or avoid litigation and should therefore be treated as privileged. A voir dire was conducted by the Tribunal on this issue and following its completion, the Tribunal dismissed the Employer's objection in a ruling rendered August 21, 1992 (see Voir Dire Ruling for further details).

46. The Employer alleges the job evaluation data generated in the course of the study is not sufficiently reliable for the adjudication of the complaints referred to the Tribunal. The Employer is not satisfied with the reliability of the evaluation results. The Employer's equalization payments indicate the extent to which the Employer is willing to rely upon the evaluation results. The Commission and the Alliance are seeking to use the evaluation data for a determination of wage disparity and pay adjustments under s. 11 of the Act.

II. ISSUE

47. As a result of a pro-active initiative by the Employer, the Complainant, together with 13 other public sector unions, and the Respondent entered into a pay equity study called the Joint Union/Management Initiative.

48. The JUMI Study began in 1985 and lasted until January, 1990, when the JUMI Study was aborted firstly by the Complainant and then by the Respondent. The Complainant and the Respondent produced, over that period of time, job evaluation results.

49. Prior to the commencement of the JUMI Study, the Complainant had filed with the Commission a s. 11 wage discrimination complaint against the Respondent. After the breakdown of the JUMI Study, the Complainant filed a second and new complaint against the Respondent.

50. The Commission and the Complainant intend to use the job evaluation results from the JUMI Study as evidence of the value of work performed by male and female employees whose jobs are the subject of these complaints. The Commission and the Complainant further intend to use the job evaluation results as proof of a wage gap alleged by these complaints as contrary to s. 11 of the Act.

51. The Respondent submits the job evaluation results are unreliable for purposes of adjudication. More specifically, the Respondent alleges the job evaluation results are biased, in as much as, the male-dominated questionnaires and the female-dominated questionnaires used to produce the results were treated differently by the individuals who performed the evaluations.

52. Therefore, the issue is whether or not the job evaluation results of the JUMI Study are reliable for purposes of the s.11 complaints referred to this Tribunal for deliberation.

III. LEGISLATION

53. The complaints before us allege wage discrimination on the basis of sex contrary to s.11 of the Act. Section 11 states:

> 11(1) It is a discriminatory practice for an employer to establish or maintain differences in wages between male and female employees

employed in the same establishment who are performing work of equal value.

(2) In assessing the value of work performed by employees employed in the same establishment, the criterion to be applied is the composite of the skill, effort and responsibility required in the performance of the work and the conditions under which the work is performed.

...

(5) For greater certainty, sex does not constitute a reasonable factor justifying a difference in wages.

54. The equal pay for work of equal value provisions of s. 11 of the Act was the subject of a Supreme Court of Canada decision in the case of Syndicat des employés de production du Québec et de l'Acadie v. Canada (Canadian Human Rights Commission), [1989] 2 R.C.S. 879 (S.C.C.). That decision dealt with the issue of whether the Canadian Human Rights Commission's decision to dismiss a complaint pursuant to s. 36(3)(b) of the Act is "required by law" to be made on a quasi-judicial basis and accordingly, reviewable by the Federal Court of Appeal under s. 28 of the Federal Court Act. The majority of the Court held that the Commission's decision was not reviewable by the Federal Court of Appeal under s. 28 of the Federal Court Act and thus, the Commission's decision was not one required to be made on a judicial or quasi-judicial basis.

55. Although the interpretation of s. 11 of the Act was not integral to the majority decision, Sopinka J. in delivering for the majority said at p. 903:

The intention of s.11 is to prohibit discrimination by an employer between "male and female employees" who perform work of equal value and not to guarantee to individual employees equal pay for work of equal value irrespective of sex.

56. In our view, as expressed by Sopinka J. the wording of s. 11 prohibits any practice by an employer to differentiate on the basis of "sex" when determining the wages or compensation to be paid between its male and female employees who perform work of equal value. For greater certainty, s. 11(5) makes it clear that "sex" does not constitute a reasonable factor justifying a difference in wages. Other sections of the Act also refer to prohibitions on the basis of sex. Section 3(1) of the

Act includes "sex" as one of the prohibited grounds of discrimination. Section 7 of the Act declares that it is a discriminatory practice to refuse employment or differentiate adversely during the course of employment on a prohibitive ground, i.e., sex. Section 10 of the Act declares that it is a discriminatory practice to establish or pursue a policy or practice or to enter into an agreement affecting recruitment, referral, hiring, promotion, training, apprenticeship, transfer or any other matter relating to employment or prospective employment that deprives or tends to deprive an individual or class of individuals of any employment opportunity on a prohibitive ground of discrimination, i.e., sex.

57. The discriminatory practice alleged in the complaints before the Tribunal is that the Employer maintains a difference in wages between male and female employees employed in the same establishment who are performing work of equal value, contrary to s. 11. There are certain exceptions to the statutory prohibition against wage discrimination as stated by s. 11(4) of the Act. That section reads:

> 11(4) Notwithstanding subsection (1), it is not a discriminatory practice to pay to male and female employees different wages if the difference is based on a factor prescribed by guidelines, issued by the Canadian Human Rights Commission pursuant to subsection 27(2), to be a reasonable factor that justifies the difference.

58. The brief legislative history of s. 11 finds that the Government of Canada declared in 1976 that it would introduce a human rights bill. The major effect of the bill would be to prohibit discrimination on the grounds of race, colour, national or ethnic origin, religion, age, sex, marital status or physical handicap. In particular, the bill would establish the principle of equal compensation for work of equal value performed by persons of either sex. (Exhibit PIPSC-82).

59. The "Background Notes" to the Canadian Human Rights Bill, issued by the then Minister of Justice, indicate that the bill would consider, in relation to a prohibited ground, discriminatory practices such as the differentiation in wages based on sex between workers performing work of equal value. The notes state, at p. 4:

> This provision is designed primarily to cope with female `work ghettoes'; it would enable workers performing one sort of job, such as secretarial work, to have their compensation related not

> only to that of other secretaries, but also to other jobs of equal value in the firm.
>
> (Exhibit PIPSC-82, p. 3)

60. In 1977, the Government of Canada enacted the Act. The intent of s. 11 of the Act is to ensure that men and women who perform work of equal value receive equal compensation. Section 11 came into force on March 1, 1978. Section 27(2) of the Act authorizes the Canadian Human Rights Commission to pass guidelines "interpreting the provisions of the Act." Since the proclamation of the Act in 1978, the Guidelines were twice promulgated by the Commission. The first set of Guidelines passed pursuant to the Act were prepared to assist in the interpretation of s. 11 of the Act and were issued on September 18, 1978. These were revoked by the Guidelines dated November 18, 1986, and gazetted in December, 1986.

61. The 1986 Guidelines describe the manner in which s. 11 of the Act is to be applied and the factors that are considered reasonable to justify a difference in wages between males and females performing work of equal value in the same establishment. The 1986 Guidelines prescribed ten factors justifying a pay differential between male and female employees performing work of equal value. None of these exceptions play a role in these complaints.

62. The dissenting opinion in Syndicat, supra, is helpful because it does address some of the prerequisite elements necessary to build a case under s. 11 of the Act. The dissent was delivered by L'Heureux-Dubé J. in which her Ladyship refers to earlier decisions of the Supreme Court of Canada, namely, Robichaud v. Canada (Treasury Board), [1987] 2 S.C.R. 84 and Canadian National Railway Co. v. Canada (Canadian Human Rights Commission), [1987] 1 S.C.R. 1114 (sub nom: Action Travail des Femmes) which reviewed complaints based on ss. 7 and 10 of the Act respectively. Both decisions make clear statements that intent is not a precondition to a finding of adverse discrimination under the Act. L'Heureux-Dubé J. notes the scope of protection under s. 11 differs from ss. 7 and 10 and says at p. 925:

> As intent is not a prerequisite element of adverse discrimination, a complainant may build his or her case under ss. 7 and 10 by presenting evidence of the type adduced by the complainant in the present case. Statistical evidence of professional segregation is a most precious tool in uncovering adverse discrimination. Section 11, however, differs from ss. 7 and 10. Its scope of protection is delineated by the concept of "equal value". That provision does not prevent the employer from remunerating differently jobs which are not "equal" in value. Wage

discrimination, in the context of that specific provision, is premised on the equal worth of the work performed by men and women in the same establishment. Accordingly, to be successful, a claim

brought under s. 11 must establish the equality of the work for which a discriminatory wage differential is alleged.

63. L'Heureux Dubé J. is of the opinion that a complainant may build a case under ss. 7 and 10 without presenting or including as part of its case the element of intent. In her Ladyship's words, statistical evidence is "a most precious tool in uncovering adverse discrimination."

64. L'Heureux-Dubé J. asserts that although the principle of "equal pay for work of equal value" is expressed in a straight forward manner, its application under s. 11 of the Act raises considerable difficulties. She maintains the concept is simple only in appearance. One element of difficulty is the concept of equality which, in her view, should not receive a technical or restrictive interpretation. In referring to the concept of equality, L'Heureux-Dubé J. says at pp. 926-27:

> The prohibition against wage discrimination is part of a broader legislative scheme designed to eradicate all discriminatory practices and to promote equality in employment. In this larger context s.11 addresses the problem of the undervaluing of work performed by women. As this objective transcends the obvious prohibition against paying lower wages for strictly identical work, the notion of equality in s. 11 should not receive a technical or restrictive interpretation.

65. Another such difficulty, according to L'Heureux-Dubé J., persists in the concept of value. At p. 928, she states:

> The notions of `skill', `effort', `responsibility' and `conditions' which one finds in the Act and the companion Equal Wage Guidelines are terms of art. They refer to the areas traditionally measured by industrial job evaluation plans.

66. Section 11(2) defines, in general terms, the manner in which the value of the work is to be assessed and establishes four criteria, namely, skill, effort, responsibility and working conditions. The criteria are defined in greater detail in s. 3 of the Guidelines, the companion to s. 11.

67. Madame Justice L'Heureux-Dubé observes that it is more than coincidence that the same four words used in this legislation were also used in the American counterpart and that these words are an indication that job evaluation plans can be used to determine whether jobs are of equal value under s. 11. However, she is of the opinion that the use of a job evaluation plan is not necessarily the only approach to the implementation of the provisions of s. 11. It is the Commission's view, as expressed in evidence by Durber, that in a s. 11 complaint, equality of work can be established through the use of a job evaluation plan but, may also be established through other less formal methodologies.

68. The Tribunal heard expert evidence that the purpose of a job evaluation plan is, in the context of a s. 11 complaint, to determine the relative worth of jobs within an organization. It involves a systematic

process which first defines and establishes factors which relate to the four criteria identified in s. 11(2) of the Act. The factors are weighted against each other for their relative importance. Each job is assessed against each factor to develop a hierarchy of jobs. Various steps or stages are involved before a hierarchy is developed which include gathering job information, defining the jobs considered for evaluation, evaluating each job and assigning scores for each compensable factor.

69. L'Heureux-Dubé J. commented on the use of job evaluation plans and the number of steps involved at p. 931:

> All steps of such a job evaluation plan involve a measure of subjectivity. Social beliefs which have traditionally led to the undervaluing of women's work may bring a certain measure of bias in the design and application of these methods. To illustrate, job content information which is supplied by the employees can contain certain characteristics which, as a result of underlying values, may be overlooked in the assessment. There may be confusion between truly compensable characteristics and stereotyped notions of what are perceived to be inherent attributes of being a woman.

70. These comments were echoed by pay equity experts who testified at this hearing. While job evaluation procedures can be controlled, to a certain extent, it is still an inherently subjective process. The value assigned to each job is an expression of opinion given by individuals and is a judgment call by the evaluators. According to Willis, the pay equity

expert and consultant to the JUMI Committee, such a procedure may incorporate both random and systematic errors of judgment.

71. Willis testified that random errors are to be expected in an undertaking as large as the JUMI and can result from a lack of sufficient job information, assumptions about particular job aspects, inconsistent application of the Willis Plan (the job evaluation plan), or simply from a differing interpretation of the job information. Willis indicated that while random differences are expected and tend to cancel each other out, patterned differences are not expected and do not cancel each other out. These patterned differences, or systematic errors of judgment, according to Willis, are evidence of bias on the part of evaluators and should be avoided.

72. Weiner, one of several pay equity experts who testified before the Tribunal, referred to the wage discrimination identified in s. 11 of the Act as one type of systemic discrimination. She describes the unintentional aspect of systemic discrimination in Volume 6, at p. 875, as follows:

> Systemic discrimination is unintentional, impersonal, built into ongoing systems, often referred to as "neutral" systems, because they were never designed to discriminate.

Also, in Volume 6, at p. 877:

> Systemic discrimination operates in systems. It goes on and on and on in the policy books and no one designed them to discriminate so it become [sic] much more difficult to identify that discrimination.

73. According to Weiner, this discrimination emanates from the practices and processes of an employer relating to compensation rather than from individual actions.

74. Of significance to the interpretation of systemic discrimination is the Supreme Court of Canada decision in CN, supra. In that decision, the Court upheld an order of a Canadian Human Rights Tribunal which imposed upon the Canadian National Railway a special employment program for employment equity. In upholding the remedial order, Dickson C.J., as he then was, in referring to the proper interpretative attitude toward human rights codes and acts said at p. 1134:

> Human rights legislation is intended to give rise, amongst other things, to individual rights of vital importance, rights capable of enforcement, in the final analysis, in a court of law. I recognize that in the construction of such legislation the words of the Act must be given their plain meaning, but it is equally important that the rights enunciated be given their full recognition and effect. We should not search for ways and means to minimize those rights and to enfeeble their proper impact.

75. Dickson, C.J. elaborated on the purpose and objective of human rights legislation and on the Court's general attitude towards the interpretation of such legislation which is to give an interpretation that will advance the legislation's broad purposes. He referred to the Supreme Court's decision in Ontario Human Rights Commission v. Simpsons-Sears Ltd., [1985] 2 S.C.R. 536, which recognized that human rights legislation is directed not only at intentional discrimination but unintentional discrimination as well, and prohibits discrimination in situations of "adverse affect discrimination".

76. The Supreme Court of Canada in CN, supra, recognized systemic discrimination in the context of employment equity as distinct from equal pay for work of equal value referred to by Weiner in her discussion relating to s. 11 of the Act. The Supreme Court recognized that s. 15(1) and by extension s. 41(2)(a) of the 1976-77 Canadian Human Rights Act as amended in 1985 were designed to resolve the problem of systemic discrimination. Dickson C.J. described systemic discrimination, at p. 1139, as follows:

> In other words, systemic discrimination in an employment context is discrimination that results from the simple operation of established procedures of recruitment, hiring and promotion, none of which is necessarily designed to promote discrimination. The discrimination is then reinforced by the very exclusion of the disadvantaged group because the exclusion fosters the belief, both within and outside the group, that the exclusion is the result of "natural" forces, for example, that women "just can't do the

> job"...To combat systemic discrimination, it is essential to create a climate in which both negative practices and negative attitudes can be challenged and discouraged. The Tribunal sought to accomplish this objective through its "Special Temporary Measures" Order.

77. In his decision, Dickson C.J. emphasized that the Order of the Tribunal, under review there, was made to implement an employment equity program which was not simply compensatory but also prospective in its provisions so as to confer benefits designed to improve "employment opportunities for the affected group in the future." Further, Dickson C.J. reasoned that such a program was designed to break the continuing cycle of systemic discrimination in the employment of women. Dickson C.J. was of the opinion that the goal of the legislation, specifically with reference to s. 41(2)(a), was an attempt to eliminate the insidious barriers which would block future job applicants, that is to say women, from the unfair employment practices that their forebears had experienced as a group. It was not, on the other hand, concerned so much with compensating past victims of discrimination or providing employment opportunities previously denied to specific individuals.

78. Dickson C.J. found the goal was not to compensate past victims or even to provide new opportunities for specific individuals who had been unfairly refused jobs or promotions in the past, rather it was an attempt to ensure that in the future applicant workers from the affected groups would not face the same insidious barriers that blocked their forebears.

79. In that case, the Chief Justice agreed with McGuigan J., the dissenting member of the Federal Court of Appeal who found that s. 41(2)(a) of the Act (now s. 53(2)(a)) is designed to enable human rights tribunals to prevent future discriminatory employment practices against identifiable protected groups. The Chief Justice also reasoned in an employment equity program there simply cannot be a radical dissociation of "remedy" and "prevention". Further, he held "prevention" is a broad term and it is often necessary to refer to historical patterns of discrimination in order to design appropriate strategies for the future.

80. We find that s. 11 does not specifically recognize the phenomenon we referred to as "systemic discrimination" and is not a well-designed vehicle for breaking the cycle of discrimination. The comments of Dickson, C.J. in the CN case, supra, need to be taken in context. In that case, an Order made by a Tribunal pursuant to s. 41(2)(a), now s. 53(2)(a), requiring the Canadian National Railway to adopt a special employment equity program in relation to the affected female group who were seeking blue collar jobs, was under appeal. It arose from a complaint-based on discriminatory employment practices and was decided in 1986.

81. The description of systemic discrimination by Dickson, C.J. in the CN case, supra, is, in our view, the kind of unintentional discrimination which s. 11 was designed to eliminate.

82. According to expert opinion, systemic discrimination has no focus or origin, only that it develops over time. It is an attitudinal

phenomenon which undervalues female work and thus differentiates against an individual or group based on gender or sex. Research has documented the group of people most commonly affected by this type of discrimination are females, and their wages and salaries, relative to male wages and salaries, are lower. This kind of discrimination is rooted in attitudes, beliefs and mind sets about work traditionally performed by males and work traditionally performed by females.

83. Counsel for the Commission submitted s. 11 is part of a statutory regime which prohibits systemic discrimination on the basis of sex and the payment of different wages between groups of predominantly male and predominantly female employees performing work of equal value. Commission Counsel further submitted s. 11 is designed to remedy the historical undervaluing of female work and to address gender discrimination in pay. Counsel submits that proof of gender discrimination in pay is found if there is a wage gap between male- and female-dominated occupational groups performing work of equal value. (Volume 218, p. 28424).

84. It is important at this point to understand the meaning of "wage gap" within the context of s.11 of the Act. The Tribunal had the benefit of expert evidence from Armstrong, with expertise in job evaluation and pay equity, who described the overall wage gap between prevalent rates of pay earned by females as compared to males. Armstrong testified in order to comprehend the wage gap one must understand the underlying factors which may have contributed to it. She stated there may well be some legitimate and unchangeable factors responsible to some extent for the existence of the wage gap.

85. A wage gap is not something clearly delineated. The Tribunal recognizes that salary differentials between male and female jobs can be a function of job requirements making some jobs intrinsically more valuable to the employer than other jobs. Such differentials are in contrast to differentials which are based entirely on gender differences and it is the latter resulting wage gap which the Tribunal believes s. 11 is intended to eliminate.

86. Section 11 incorporates the concept of equal pay for work of equal value in its wording. Weiner testified there are two questions which arise when one invokes this concept in the context of evaluation of jobs employing the same criteria, firstly, the identification of what is meant

by "equal value" and, secondly, to define what is meant by "equal pay". Weiner equates the concept of "equal pay for work of equal value" with the concept of "pay equity".

87. The evidence before the Tribunal is that pay equity legislation addresses a trend that assumes systemic discrimination against female-dominated jobs. Some provinces have enacted pay equity legislation to remedy pay discrimination by identifying and redressing the wage gap through the implementation of pay equity plans. This latter legislation is "pro-active" because, Weiner says, its motive and intent is to provide a framework for redressing wage discrimination, rather than laying blame upon employers or unions for historical wage discrimination. The difference between pro-active legislation and s. 11 is that s. 11 is "complaint based

legislation", whereby a complainant alleges discrimination against some identified comparator group. Since s. 11(1) talks about discrimination between male- and female-dominated jobs either way, Weiner says presumably under s. 11 one could have a male job alleging discrimination.

88. While the principle of equal pay for work of equal value underpins the provisions of s.11 and is frequently expressed as "pay equity", there is in current usage of that phrase a pro-active connotation. There is, in fact, a significant difference between the principle enshrined in s. 11 which is complaint based and the pro-active approach to the problem of wage disparity which the experts in the field today accept and refer to as "pay equity". The comments of Weiner in her testimony before the Tribunal are instructive and illustrative of the problem which is encountered when applying the principles of the Act and in particular s. 11 to remedying systemic discrimination in the work force. She stated in Volume 16, at p. 2124:

> I agree with you that the Human Rights Commission law, including Section 11, is written with a complaint-based mind set. I think that was a mistake, but we didn't know that in 1977 when it was written. And really, while that makes a great deal of sense for many of the kinds of issues the Human Rights Act has to deal with, it doesn't fit as well with the systemic discrimination of something as complicated as the wage setting process.
>
> So I think you are right, there is, to my mind, an anomaly of law makers, including a methodology that fit our 1970s thinking of how discrimination operated with some forward thinking about another problem, but not recognizing that the equal value pay equity

problem was a systemic problem and didn't fit as well with a complaint-based mentality. [emphasis added]

89. Weiner, who is co-author of Pay Equity: Issues, Options and Experiences with Morley Gunderson, summarizes at the end of Chapter 8, at pp. 127-28, their conclusion regarding the federal legislation as follows:

> That pay equity is an idea whose time has come is demonstrated by the initiation of pay equity in eight Canadian jurisdictions since 1985. In the previous ten years, only two jurisdictions had passed pay equity legislation. Unlike most of the subsequent legislation, these two early pieces of legislation, in the federal government and in Quebec, were complaint-based. The inability of such legislation to address a systemic problem like pay equity is evidenced by the employer-initiated enforcement mechanism in most of the recent legislation.

90. Some change has been instituted through the political movement in the United States to enact comparable worth plans which, in turn, has created a framework within which previously invisible or unacknowledged skills associated historically with female and minority work were made visible and worthy of compensation. The parallel pay equity movement in Canada saw the enactment of provincial legislation designed to redress systemic wage discrimination and compensation for work performed by

employees of female-dominated jobs. Of relevance is the preamble to the Pay Equity Act (Ontario), 1987, which states that affirmative action is required to redress systemic wage discrimination. However, the legislative history of s. 11 does not document the same political motivation contained in that legislation or other provincial legislations found in Manitoba, Ontario, Prince Edward Island, Nova Scotia and New Brunswick.

91. Provincial legislation is aimed at correcting systemic discrimination and provides a time frame and a procedure for achieving pay equity. The approach in the provincial legislation is future oriented and while recognizing past injustices, the remedies are focused on achieving equity in employment as well as in pay. On the other hand, s. 11 of the Act is complaint based and is silent on the means for achieving equal pay for work of equal value. While the Guidelines passed pursuant to the Act expand on the four essential elements of s. 11(2), i.e., skill, effort, responsibility and working conditions and define how value is to be assessed, who is an employee, what a group is and so on, it does not establish a programme or describe an appropriate methodology for achieving

the goal of eliminating "systemic discrimination". It is a phenomenon which is not expressly referred to either in the Act or in the Guidelines.

92. Referring again to Weiner commenting on s. 11, she states in Volume 16 at p. 2125, "the legislation does not recognize that equal pay for work of equal value was a systemic problem that didn't fit as well with a complaint based mentality." Therein lies the difficulty with s. 11 which is not entirely compatible with the evolution and application of the principles of pay equity (or comparable worth) during the past two decades. Nevertheless, it is necessary in view of the general nature and intent of the legislation which is to combat systemic discrimination to adopt the reasoning of Chief Justice Dickson, at p. 1139 of CN, supra, where he states:

> ...it is essential to create a climate in which both negative practices and negative attitudes can be challenged and discouraged.

93. The wage setting process in the Federal Public Service is a highly complex process spanning many decades, each contributing to new trends and developments, most notably, the introduction of collective bargaining in the 1960s. The advent of collective bargaining brought about contract negotiations which in turn affect the determination of wage rates. For the most part, job classification in the Federal Public Service has been determined by a job evaluation process; however, no single process has ever been stipulated and the result is a classification structure of multiple occupational groups with no common job evaluation plan. Rates of pay have been arrived at through this process with the aid of labour market surveys, largely provided by the Pay Research Bureau until 1992. It is apparent the classification system has been undergoing reform since 1990.

94. Evidence was led that the Government of Canada is committed to simplifying the job classification system in the Public Service through an initiative entitled PS2000. Part of this Initiative is a commitment to compensate employees equitably, in a manner that is free of gender bias and

maintains equal pay for work of equal value. A new classification system is being introduced to meet these commitments.

95. Documentary evidence reveals a Task Force has been examining and developing this initiative and has thus far produced a draft pamphlet in November, 1992 as a reference guide for public sector employees to prepare what is referred to as "gender-neutral" work descriptions

96. The expert evidence reveals that compensation systems that rely on market surveys can result in wage disparities for jobs deemed to be of equal value. Research has shown that the market reflects an historical pattern of lower wages to employees in positions staffed predominantly by females. For the most part, market rates are established through the use of traditional job evaluation systems which self-perpetuate the problem of undervaluation of female work as these traditional job evaluation systems were not designed to capture skills associated with female work.

97. The pay equity experts explain that gender bias is reflected in existing compensation systems and pay practices. Historically these systems and practices undervalue female work. Since the purpose of s. 11 is to remove gender discrimination from pay, based on the intrinsic value of a job, any job evaluation system used to assess job value must be designed to eliminate factors that contribute to gender bias and include factors that will capture skills associated with female work which have, in the past, been overlooked.

98. We do find that s. 11 is a remedial section dealing with salary inequities which arise between jobs that are deemed by some process of evaluation to be of equal value. The salary inequity, or resulting wage gap, is the salary differential between rates of pay for male and female employees who are performing work of equal value and not the overall wage gap referred to by Armstrong. She was referring generally to differences in pay between males and females which can result from factors in addition to gender inequities. Armstrong's response to the following question in Volume 179, at p. 22879, lines 2 - 7, is informative:

> If pay equity were to be achieved in all occupations, in all jobs, would the wage gap disappear?
>
> THE WITNESS. The overall wage gap probably wouldn't disappear completely, no. There still might be a difference.

99. We must be assured the complaints seek to redress a wage gap based on wage differentials that are gender based and not resulting from other factors. It seems apparent that the existence of a wage gap per se is not proof of discrimination. To hold otherwise would negate the entire evaluation process which has, as its purpose, the comparison of jobs according to a plan or system for rating work according to the criteria prescribed in s. 11(1) of the Act.

100. We also find s. 11 is designed to eliminate economic inequality created by gender based wage discrimination. The discrimination is unintentional as the decision of Dickson C.J. in the CN case, supra, makes

clear. It is nevertheless a subtle form of discrimination built into employment practices as they have existed over the years since females have become contributors to the work force. We recognize from the expert testimony of Weiner, Armstrong and Willis that systemic discrimination operates in systems and becomes incorporated into the wage setting practices of organization and that classification of jobs may be the by-product of systemic discrimination. Since systemic discrimination is part of a system never designed to discriminate, Weiner says that it cannot be corrected instantly nor can pay equity be achieved quickly.

101. As remedial legislation, s. 11 addresses pay disparities in employers' compensation practices. Willis testified to the effect s. 11 is not true pay equity legislation, but instead concerns itself with examining pay disparities and is probably a first step in the direction of true pay equity insofar as it requires the wages of females be moved up to the same level as males. Willis testified in Volume 29, at p. 3760, line 22 to p. 3761, line 9:

> The concept of pay equity has to do with compensation without gender bias; that is, compensation based on the intrinsic value of a job rather than the market value of a job.
>
> I recognize there is a school of thought that says: Don't bother with job evaluation, just give us the money.
>
> But in order to logically arrive at an intrinsic value of a job, you exercise a job evaluation plan; that is, job evaluation provides a way of taking any job apart and examining the amount of skill, effort, responsibility, and conditions of work that are required.

102. If employers use job evaluation systems that are gender biased in favour of male work, the result will be seen in differential wages paid to male and female jobs that ought to be considered of equal value. Job evaluation systems that traditionally favour male jobs do not value the skills and job content of jobs that are designated female work. Traditional job evaluation is most often designed to value characteristics of male work. On the other hand, pay equity job evaluation has as its goal the use of systems that remove gender bias in the valuing of work.

103. At this point, it is useful to recall some of the circumstances leading up to the issue before us. Under the JUMI, the parties engaged in a proactive study with the intention of developing parameters in which to

implement the principal of equal pay for work of equal value as incorporated in s. 11 of the Act. The parties employed a pay equity expert, Willis, to assist with this study and used the Willis Job Evaluation Plan to assess selected jobs from male- and female-dominated occupational groups in the Federal Public Service. The JUMI never completed its task. The data generated in that study is now in evidence before the Tribunal. It was used by the Commission in its investigation of the complaints and is presented as proof of a breach of s. 11 of the Act. The Commission and the Alliance have called upon the Tribunal to accept the evaluation scores as evidence of the value of work. These results, they

submit, can be used to establish the equality of work and are proof of a wage gap. It is alleged by the Employer that the results are not reliable. We are called upon to determine whether these results are reliable.

104. We have previously referred to the steps involved in the use of the job evaluation plan as discussed by L'Heureux-Dubé J. in the Syndicat decision, supra. The Tribunal heard lengthy evidence on the Willis Process which incorporates the typical steps involved in job evaluation. The reliability of the results, which is the issue before us, focuses on one particular important step in job evaluation, namely, its application by the evaluators who had the responsibility of analyzing the job information and assigning points or scores to each of the job factors in the Willis evaluation plan.

105. Willis and Weiner agree for job evaluation to be effective in eliminating bias it must be approached in a systematic fashion. At the same time, one needs to understand that job evaluation is an inherently subjective process. The question as to what constitutes "bias" is a complex one, and is central to the arguments submitted to us. Counsel for the Commission refers to the Equal Wages Guidelines passed pursuant to the provisions of the Act and in particular to s. 9(a) which reads:

> 9. Where an employer relies on a system in assessing the value of work performed by employees employed in the same establishment, that system shall be used in the investigation of any complaint alleging a difference in wages, if that system
>
> (a) operates without any sexual bias;

106. Commission Counsel, supported by the Alliance, posits the question of the reliability of the results to be addressed by the Tribunal as follows:

Is there a pattern, a systematic variance of different treatment of male and female questionnaires (in the evaluation process) that was caused by or is attributed to gender bias or gender related bias. [emphasis added]

107. Respondent Counsel advocates a broad reading of the term "sexual bias" as used in s. 9(a) of the Guidelines. Respondent Counsel proposes that any bias that is a different treatment of male and female questionnaires is a sexual bias and bases this submission on an interpretation of Willis' testimony, who described bias in Volume 208, at p. 26937, lines 11 - 16, as follows:

> A. "Bias" simply means that if there is a pattern of different treatment for male-dominated jobs versus female-dominated jobs, whether it's conscious or unconscious, that difference in treatment would represent an amount of bias.

108. According to Willis, it is possible to have a bias that is related to gender but is not a direct gender bias, and he refers to this bias as a gender preference. His example of a gender preference would be

where an evaluator may have a preference for individuals who wear blue shirts or have blue collars. Willis testified, for example, trade jobs are known as "blue collar" jobs. If a preference for blue collars causes an evaluator to evaluate trades jobs more favourably, Willis says that this may not be a gender bias but maybe it is a blue collar bias. According to Willis, this would bring the same result as a direct gender bias.

109. The Employer submits that the meaning of gender bias can include attitudes toward one sex or the other that are conscious or unconscious. In the Employer's view, a bias can also relate to some characteristic which is not gender per se, but is itself related to gender, which they describe as a "gender-related bias". Respondent Counsel submits that s. 11 is designed to redress both kinds of biases. Respondent Counsel also submits that if a "blue collar" preference results in a different treatment of male and female questionnaires this is a "sexual bias" as contemplated by s. 9(a) of the Guidelines. Respondent Counsel urges that the question to be addressed by the Tribunal is not the one posed by the Commission but rather is as follows:

Is there a pattern of different treatment of male and female questionnaires?

110. It is to be noted that the Commission and the Alliance do not express any difficulty in assigning a wide meaning to the term "sexual bias" and we refer to the remarks of Commission Counsel in Volume 230, at p. 30583, lines 14 - 25:

> These are very different things. All those other things such as -- I mean, they referred to something called "dirty work'". I don't know whether you would have a preference for people who do hard work outdoors. If that were gender related and had an effect on the way people perceive the work and rated the jobs, and in the end had a consequential gender effect on the jobs, then no one can dispute that that would be a gender bias contrary to section 9 of the Equal Wage Guidelines and therefore contrary to section 11 of the statute.

111. In formulating the question for the Tribunal to address, Respondent Counsel argues that their formulation does not require a causative factor for the different treatment of male and female questionnaires. The disagreement between the parties lies not in assigning a broad meaning to the words "sexual bias" but instead arises as to whether s. 11 requires the existence of a cause when different treatment of male and female questionnaires is found or whether, on the other hand, it is simply a matter of differential treatment of male and female jobs without the necessity of assigning cause. In support of the Employer's submission, they rely on a meaning of bias which in their view does not require a causal link or relationship under s. 11 of the Act.

112. There is a disagreement between the parties about the analysis and investigative findings of the Commission on the job evaluation process and the statistical evidence. The dispute centres on the submissions of the Commission and the Alliance that some differences in treatment of male

and female questionnaires between committees and consultants are not based on gender or gender-related bias but are due to a "value bias". The Commission and the Alliance rely on the statistical expert, Sunter, whose analyses, they submit, demonstrates the effect of a value bias which accounts for some, if not all, of the differences in treatment between the committee and the consultant. Sunter testified that the effect of the value bias has an appearance of gender bias and the difference in treatment between the committees and the consultants is as likely to be a consequence of value bias as it is gender bias.

113. In dismissing the need to know the cause of differential treatment of male and female questionnaires, Respondent Counsel relies on the testimony of Willis. Willis repeatedly stated during this hearing that after the evaluation process is finished there is no need to explore the reasons for the differences between the evaluations of the committees and the consultants. That testimony can be summarized in Willis' letter to Respondent Counsel dated May 19, 1994 which expands on job evaluation disparities. This reads as follows:

> Evaluation disparities represent a lack of consistency in the application of the evaluation system. Therefore, disparities are a cause for concern, and require attention to determine if they result in a pattern of different treatment for different kinds of jobs.
>
> The question as to why disparities have occurred is important during the course of the committees' work. An understanding of the reasons can be helpful in the continued training of the members. However, after the evaluation phase of the study has been completed, the reasons for any disparities are no longer of any real importance. What is important is the existence of any pattern of bias that is developed among the evaluations. [emphasis added]

(Exhibit R-164)

114. In the course of our hearing, in addition to his definition of bias as a different treatment between male and female jobs, Willis offered an opinion on the meaning of gender bias in a pay equity study in Volume 80. He states in Volume 80, at p. 9737, lines 13 - 18:

> In the context of the pay equity study, gender bias has to do with the extent to which jobs that are traditionally held by one sex or the other are paid more favourably than jobs that are traditionally held by the opposite sex.

115. Willis refers to gender bias as both different treatment and different pay. To better understand Willis' definitions of bias, it is helpful to refer to the theory of disparate treatment considered in the decision American Federation of State, County and Municipal Employees, ASL-CIO et al. v. State of Washington et al., Nos. 84-3569, 84-3590, 770 s. 2d 1401 (1985) United States Courts of Appeal, 9th Circuit.

116. The plaintiffs in the American case alleged sex discrimination in compensation against the State of Washington pursuant to s. 703(a) of Title VII of the Civil Rights Act of 1964, 42 U.S.C. The United States District Court for the Western District of Washington had found in favour of the class of state employees of which at least 70 per cent were female, and the state had appealed to the Court of Appeal, 9th Circuit. A relevant fact in the District Court decision, was that Willis had conducted a study in 1974 to examine and identify salary differences pertaining to job classes predominantly filled by males compared to job classes predominantly filled by females, based on job worth. The 1974 Willis Report submitted into evidence concluded based on the job content of the 121 classifications evaluated, the tendency was for female classes to be paid less than male classes for comparable job worth, and that overall the disparity was approximately 20 per cent. Willis' study had deemed the male and female positions to be of comparable worth. Comparable worth as defined by the State, for the District Court, means the provision of similar salaries for positions that require or impose similar responsibilities, judgments, and knowledge.

117. In the first instance, the district court had found a violation of Title VII premised upon the American disparate impact and the disparate treatment theories of discrimination. As explained in the District Court's judgment, Title VII prohibits two types of employment discrimination: (i) intentional unfavourable treatment of employees based upon impermissible criteria; and (ii) practices with a discriminatory impact: facially neutral practices that have a discriminatory impact and are not justified by business necessity.

118. The District Court decision was appealed to Kennedy, Circuit Judge for the United States Court of Appeals, who considered the allegations of disparate treatment, and held that the unions had failed to prove a prima facie case of sex discrimination by the preponderance of the evidence. In citing reasons, Kennedy J. offers the following with regard to the Willis study at p. 1408:

> We also reject ASFCME's contention that, having commissioned the Willis study, the State of Washington was committed to implement a new system of compensation based on comparable worth as defined by the study. Whether comparable worth is a feasible approach to employee compensation is a matter of debate...Assuming, however, that like other job evaluation studies it may be useful as a diagnostic tool, we reject a rule that would penalize rather than commend employers for their effort and innovation in undertaking such a study.

119. As noted in the decision of Kennedy J., under the disparate treatment theory, an employer's intent or motive in adopting a challenged policy is an essential element of liability for violation of Title VII. To establish liability, a plaintiff must show the employer chose a particular policy because of its affect on members of a protected class and it is insufficient for a plaintiff to allege under this theory that the employer was merely aware of the adverse consequences the policy would have on a protected class.

120. The United States Court of Appeals had to find a proof of intent required in a disparate treatment case unlike s. 11 which addresses systemic discrimination, a form of unintentional discrimination. Willis' definitions of bias should be viewed within the context of that American jurisprudence which we note deals with a different statute than the Act and a different requirement of intention.

121. Referring to s. 9(a) of the Guidelines, supra, we note that it provides, inter alia, that an employer may use a system for assessing work if that system operates without any sexual bias. By way of contrast and focusing on the issue which we must resolve, it is in the application of the system that we are concerned. In this regard, it is helpful to refer to the comments of Weiner, and to the following statement she made:

> Even though I mentioned gender bias in job evaluation and gender bias in the evaluation systems, I think gender bias in the application of the system is key. If you give people bias free job information and a bias free evaluation system, people can still introduce gender bias when they apply it.

122. There is agreement between the parties that the evaluation system, i.e., the Willis Job Evaluation Plan, is bias free by any reasonable standard. According to Willis, if there is a pattern of differential treatment between male and female questionnaires, this is evidence of systematic biases occurring in the application of job evaluation. For purposes of determining whether bias is present in the results, Willis was not prepared to give an opinion based solely on his observations of the committee process, but instead said he would rely on statistical analysis of the data. According to Willis, there are two ways to determine if bias is present in the application of the plan: (i) observation by the consultants who participated in the process; and (ii) statistical analysis.

123. It was Willis' opinion that as a factual matter, he and his consultants who were present during the job evaluations, were able to ferret out and identify direct gender bias. They observed how evaluators responded to questions as to why they evaluated jobs a particular way. The consultants would not permit a rater to defend an evaluation based on opinions or conclusions.

124. Willis said that to the extent that indirect biases occur, they are more difficult to detect. Usually the only way to detect if indirect bias is operating is to do a statistical analysis of the results to determine if a pattern in the ratings exist. Since the evaluators are usually unconscious of these biases, they are not aware of making gender based judgments. Evaluators will apply points unevenly across male and female jobs, and male and female jobs will consistently receive low or high points. Generally speaking, a statistical analysis will reveal this type of pattern if indirect biases have entered the process.

125. Both the question posed by the Employer and the question of the Commission, in our view, restrict the Tribunal from fully assessing the question of reliability. The issue of reliability is not purely

statistical and the questions as suggested restrict our assessment of the evidence to statistical measures.

126. It is important to bear in mind that the results were generated through a process of job evaluation overseen by a JUMI Committee with the advice and consultation from a pay equity expert. Willis testified that he recommended to the JUMI Committee certain safeguards in the process to ensure consistent and reliable results. These safeguards included reliable job information, balanced evaluation committees, selection and rating of benchmarks, sore-thumbing exercises, training of participants, quality checks in the form of testing of evaluators and committees and consultant participation.

127. Willis took a consistent position throughout the hearing that in order to analyze the results, he required a statistician. He viewed the role of the statistician to examine the data and to ascertain the extent of the problem if one was found.

128. Willis said he would not support the reliability of the results based on the process alone. When asked to consider the process without the results, Willis said in Volume 78, at p. 9570, lines 12 -22:

A. ...if all of my recommendations had been taken, if I had felt that the processes that were followed were all sound, then it is quite likely that I would have been able to support the results of the study without doing any testing.

This didn't happen. I have not yet supported the results of the study. But in the final analysis that testing is going to tell somebody, me or somebody else, whether or not the study was sound.

129. Expert Weiner has stated that the idea behind a job evaluation process is to be systematic, that the process should involve a series of steps. The goal in pay equity job evaluation should be, in her opinion, to apply the process fairly to all the jobs. Notwithstanding Willis' opinion which focuses primarily on results rather than process, the Tribunal must be able to assess the checks and balances in the process, and must be able to do this not only from a statistical perspective but also to analyze reliability by assessing how the Willis job evaluation plan was applied by the job evaluation committees.

130. We are entitled to look at the Act as a whole, including the regulations and Guidelines passed pursuant thereto, in order to assist us in interpreting the meaning of s. 11(1), see Driedger, The Construction of Statutes, Chapter 11, 3rd Edition by Ruth Sullivan. It is our opinion the legislation, given the state of development or evolution of the concept of equal pay for work of equal value at the time, its complaint based orientation and considering the gender driven language of the relevant sections, that causation is implicit in its provisions.

131. The wage gap to be redressed by s. 11 must be caused by gender based discrimination. Section 9(a) of the Guidelines is subordinate to the enabling legislation, the Act, and is authorized by s. 27(2) of that Act.

There is a presumption in favour of the validity of regulations in the light of their enabling statute. In the Interpretation of Legislation in Canada, 2nd Edition, Pierre Andre-Coté at p. 310 the learned author comments as follows:

Finally it must be pointed out that the regulations are not only deemed to remain intra vires, but also to be formally coherent with the enabling statute.

132. Moreover, s. 16 of the Interpretation Act (Canada) provides:

> Where an enactment confers powers to make regulations, expressions used in the regulations have the same respective meanings as in the enactment conferring the power.

133. For purposes of s. 11 of the Act we do not find it necessary to make a distinction between gender bias or gender preference. We are in agreement with the parties that the phrase "sexual bias", as contained in s. 9(a) of the Guidelines, should provide for any bias in the context of job evaluation which has the end result of favouritism toward one gender. Moreover, we agree with Willis and the Employer that it is not necessary to determine why a particular evaluator is motivated to exhibit bias. However, we find it necessary to examine the differences between committees and consultants from both a statistical perspective and a process perspective to determine if a bias exists.

134. In our opinion, causation is implicit in the provision of the legislation and the Guidelines. Different treatment of male and female jobs must be proven to be gender-based. This is consistent with the opinions expressed by Willis as he does not merely talk about a different treatment but a different treatment that is "an influence towards one gender or another" (Volume 38, p. 4794) and a bias "favouring" a gender (Volume 38, p. 4792). It is the gender aspect of the treatment that concerns Willis and which concerns the Tribunal.

135. Accordingly, the Tribunal is interested in the gender aspect and based on our interpretation of the Act, the question to be addressed is:

Is there a different treatment of male and female questionnaires in the evaluation process that was caused by or attributed to gender bias or gender-related bias?

136. We will now address the question of whether there is gender-based bias present in the treatment of male and female questionnaires. Our enquiry will encompass the evidence of the process that generated these results and the statistical evidence presented at this hearing.

IV. BURDEN OF PROOF

137. In the case before the Tribunal the issues, because of the length and complexity of the evidence, have been argued and addressed by the parties in stages. The first stage relates to the reliability of the results generated by the JUMI study. The affirmative alleges that the

results are reliable and free of gender bias by any reasonable standard. The negative alleges that the results are unreliable and coloured by gender bias to such a degree that they do not allow for an adjudicated resolution by the Tribunal.

138. The phrase "burden of proof" describes the duty which lies on one or the other of the parties either to establish a case or to establish the facts on a particular issue. See M.N. Howard, ed. Phipson on Evidence, 14th ed., (London: Sweet & Maxwell, 1990) para. 4-01.

139. In Miller v. Minister of Pensions, [1947] 2 All E.R. 372, (K.B.), Lord Denning, at p. 374, defines the degree of probability required to discharge the burden of proof in a civil case in these terms:

> That degree is well settled. It must carry a reasonable degree of probability but not so high as is required in a criminal case. If the evidence is such that the tribunal can say: "We think it more probable than not" the burden is discharged, but if the probabilities are equal it is not.

140. Counsel for the Commission in her opening remarks conceded that the burden of establishing a prima facie case rests with the Alliance and the Commission. (Volume 218, p. 28337).

141. This concession by Counsel for the Commission simply recognizes the evidentiary rule frequently enunciated by the Courts and contained in the text books on this subject.

142. In the view of Sopinka J. et al, The Law of Evidence in Canada, (Toronto: Butterworths, 1992), a prima facie case does not compel a specific determination unless there is a specific rule of law which demands such a conclusion. After examining and analyzing several decisions of the Supreme Court of Canada in which the Justices differ, the authors state that a prima facie case simply permits an adverse finding against the Employer in the absence of evidence to the contrary. The authors quote with approval a passage found in R. v. Girvin (1911), 45 S.C.R., 167, (S.C.C.) at p. 169 as follows:

> I have always understood the rule to be that the Crown, in a criminal case, is not required to do more than produce evidence which, if unanswered, and believed, is sufficient to raise a prima facie case upon which jury might be justified in finding a verdict." [emphasis added]

143. This passage was recently adopted in R. v. Mezzo, [1986] 1 S.C.R. 802 (S.C.C.) and the learned authors conclude at p. 73:

> The terms "prima facie evidence", "prima facie proof", and "prima facie case" are meaningless unless the writer explains the sense in which the terms are used. For clarity and conciseness it is preferable...to explain the evidentiary effect consequent upon the proof of certain facts rather than to indiscriminately use these mixed Latin English idioms.

30

144. Because there appears to be some question as to the meaning of the phrase "burden of proof" as it applies in these circumstances we refer again to Phipson on Evidence, supra. According to the learned author, it has three meanings as follows:

(i) The persuasive burden, the burden of proof as a matter of law, i.e., the burden of establishing a case by a preponderance of evidence;

(ii) The evidential burden, the burden of adducing evidence; and

(iii)The burden of establishing the admissibility of evidence.

145. The persuasive burden, sometimes referred to as the "legal burden", in a civil case rests on the party who substantially asserts the affirmative of the issue and is fixed at the beginning of the trial or hearing by the state of the pleadings, i.e., the complaints made pursuant to the legislation, and it is settled as a question of law that the burden remains unchanged throughout the hearing exactly where the complaints place it, and only rarely shifts except under special circumstances.

146. The legal burden of proof normally arises after the evidence has been completed and the question is whether the trier of fact has been persuaded with respect to the issue or case to the civil or criminal standard of proof. The legal burden, however, ordinarily arises after a party has first satisfied an evidential burden in relation to that fact or issue. See The Law of Evidence in Canada, supra, at p. 58.

147. Stated another way, the legal burden does not play a part in the decision making process if the trier can come to a determinate conclusion on the evidence. If, however, the evidence leaves the trier in a state of uncertainty, the legal burden is applied to determine the outcome on a balance of probabilities. See also The Law of Evidence in Canada, supra, at p. 60 quoting a passage in a decision of the Privy Council in Robins v. National Trust Company, [1927] 2 D.L.R. 97, which reads in part as follows:

> But onus as a determining factor of the whole case can only arise if the tribunal finds the evidence pro and con so evenly balanced that it can come to no sure conclusion. Then the onus will determine the matter.

148. This passage can be compared to the comments of McIntyre J. in Ontario Human Rights Commission v. Simpsons-Sears [1985] 2 S.C.R. 536 at 558:

> But as a practical expedient it has been found necessary, in order to insure a clear result in any judicial proceeding, to have available as a `tie breaker' the concept of the onus of proof.

149. The evidential burden, on the other hand, may shift constantly throughout the hearing, accordingly as one scale of evidence or other

preponderates. The burden of proof in this sense rests upon the party who would fail if no evidence were produced at all, or no more evidence, as the case may be, were given on either side. In civil cases the evidential burden may be satisfied by any species of evidence sufficient to raise a prima facie case. It is for the Tribunal to decide as a matter of law whether there is sufficient evidence to satisfy the evidential burden, that is to say, to establish a prima facie case. See Phipson on Evidence, supra, at para. 4-10(b).

150. The burden of proof in any particular case depends on the circumstances in which the claim arises. In general, according to Phipson, the rule which applies is "he who invokes the aid of the law should be the first to prove his case." This rule is founded on considerations of good sense and as well, in the nature of things, a negative is more difficult to establish than an affirmative. See Robins v. National Trust Co., supra,; Constantine Line v. Imperial Smelting Corp., [1942] A.C. 154, 174 per Lord Maugham.

151. Commission Counsel, in her oral presentation, in Volume 218, at p. 28349, line 25 to p. 28350, line 11, asserts as follows:

> If a process is created which is considered by the experts to be the best process for identifying gender bias...then there is no reason to look further beyond that process. If that's the case, then there is prima facie evidence of a reliable process, which is [sic] the absence of evidence to the contrary would permit a finding of reliability. [emphasis added]

152. This rather broad statement of Counsel is supported by reference to Farnquist v. Blackett-Galway Insurance Ltd. (1969), 72 W.W.R. 161 (Alta. C.A.) (Allen J.A.) at pp. 172-73 and by OPSEU v. Ontario (Ministry of Community and Social Services) (1986), 15 O.A.C. 78 (Div. Ct.) at p. 79 which deal with proof on a balance of probabilities.

153. It is not clear what Counsel intends by use of the phrase "If a process is created". For purposes of clarification, if Counsel means the procedures and structures put in place by Willis and the JUMI Committee for the evaluation of jobs, the evidence establishes that in general the structures are compatible with the requirements of the Act, the Guidelines and with the principles of pay equity as understood by the experts.

154. Assuming the process encompasses, not only, the procedures and structures such as the evaluation plan, the training of the evaluators, the questionnaires, the collection of information and in addition but more importantly, according to Weiner, the application of the evaluation system in a gender free manner, then one might accept Commission Counsel's statement as correct. But Counsel follows her statement with this comment in Volume 218, at p. 28357, line 22 to p. 28358, line 7:

> The purpose of all this is to relate to the shifting burden that rests with the Defendant, once the Complainants have demonstrated that there is a prima facie case and the burden has shifted to the Defendant, it is incumbent upon them to prove on the balance of

> probabilities...that gender bias or whatever their allegation is going to be is in fact the cause of the event and, therefore, is in fact the cause of the unreliability of the results. [emphasis added]

155. The shifting burden of proof referred to by Counsel in the passage quoted above does not relieve the party who asserts the affirmative, in this case the Commission and the Alliance, from satisfying the evidential burden that the results of the study can and ought to be relied upon for purposes of adjudication. If, on the whole of the evidence, including both anecdotal and statistical testimony, the Tribunal can come to a determinate conclusion it will not be necessary, in our opinion, to invoke the legal burden, in order to reach a decision.

156. Counsel for the Commission in his opening remarks admitted that "the process" in the JUMI exercise was flawed but not so flawed as to vitiate the results. The anecdotal testimony of the participants in the

study, the Willis consultants and some of the evaluators, raise questions about the impartiality of some evaluators as well as the functioning of certain committees. Incidents occurred which were disturbing and caused the consultants to experience discomfort about the process and its application. Additionally, there were differences between the committees and the consultants in re-evaluation exercises conducted by the consultants at different stages of the study. Analyses of the results, in turn, led to a critique of the data by qualified statistical experts. This short description of some of the problems which arose during the course of the study and which might have had some effect on the scores of the evaluators is not exhaustive.

157. The problems relating to the reliability of the results whether arising from the evaluation process or from the statistical analyses on re-evaluations will be addressed in a subsequent portion of this decision.

158. The Employer submits if the Willis Process worked well then the Complainant and the Commission have made out a prima facie case on reliability and there is no need therefore to look at the results for further evidence of reliability. If on the other hand, the process did not work well, then the onus, according to the Employer, remains with the Alliance and the Commission to demonstrate through "other evidence" that the results are reliable and sound.

159. What is meant by "other evidence" is described by Respondent Counsel as consisting of statistical analyses performed by the statistical experts to demonstrate there is a systematic pattern in the disparities. But Counsel argues that by "attacking the credibility and the usefulness of those disparities" the Commission and more particularly the Alliance are left with no basis for comparison between the committee evaluations and the re-evaluations performed by the consultants, whose credibility and impartiality were under attack by the Alliance.

160. According to Respondent Counsel, in the eventuality that the process did not work well and if the Commission and the Alliance are to be

precluded from relying on the statistical analyses, they are left with nothing and have therefore failed to establish a prima facie case.

161. With respect, the Tribunal, is unable to accept the proposition advanced by Respondent Counsel. In our view, it is not simply a matter of choosing to accept or reject one or both of the alternatives presented to us.

162. Within the JUMI Study itself and under the direction of Willis, the approach adopted by the JUMI Committee was to use statistical tests as a means of validating the process. The whole Willis Process is a complex scheme that does not only include an exercise of job evaluation but is inclusive of many steps and stages, one of which is the validation of the results by testing for inter-rater and inter-committee reliability. Some of the testing uses statistical analysis. These tests are integral to the process Willis utilizes in large pay equity studies, and are most evident in the JUMI Study. Other significant tests undertaken were a re-evaluation of 222 positions by the Willis consultant, Jay Wisner, who performed statistical tests on the re-evaluation results. The only statistical testing which did not occur during the process itself consisted of an additional re-evaluation of 300 positions conducted by the Commission in its investigation after the completion of the study.

163. The statistical analysis by the Commission combined the re-evaluations which occurred during the JUMI Study with the re-evaluations that were done subsequently. The act of combining these re-evaluations does not, in our view, create any artificial framework in respect of the evidence as it relates to the process or the evidence as it relates to statistical analyses. Neither process nor statistical measures operated in complete isolation from each other, but were interlocked in the sense that an understanding of one required an understanding the other.

164. Accordingly, we are entitled to look at the whole of the evidence and to weigh it in the light of all the circumstances. We will be examining in great detail the testimony of the participants in the study, the expert evidence of the consultants, the expert testimony of the statisticians and others who had some involvement in the study. Our decision, will therefore encompass all the evidence presented to us during the course of this hearing.

165. Those elements required to satisfy the evidential burden in the present proceedings consist, in our opinion, of the following which are based on the provisions of s. 11 of the Act, the companion Guidelines and the state of the pleadings:

(i) The complainant groups are female-dominated within the meaning of the Equal Wages Guidelines;

(ii) The comparator groups are male-dominated within the meaning of the Equal Wages Guidelines;

(iii)The value of work assessed is reliable; and

(iv) A comparison of the wages paid for work of equal value produces a wage gap.

166. As mentioned previously at this stage the Tribunal will address the third element which is have the sampled positions in the JUMI Study been properly evaluated so as to produce reliable results. It should be noted moreover, the parties, including the Employer, have agreed the Willis Plan is, in fact, an appropriate gender free evaluation plan for the JUMI Study which captures the criteria required to be measured by s. 11(2) of the Act.

167. In addressing the third element relating to the reliability of the evaluations, Counsel for the Commission enumerated several considerations which needed to be taken into account, namely: the plan allows for comparison between occupations; the process was designed to obtain reasonably reliable job information; there were additional procedures in place so as to ensure comprehensive job information; the Plan was, in fact, applied with reasonable consistency by the multiple committees; there was consistency in the job information; there was consistency in the results; and the salary data was reasonably reliable.

168. These considerations are, it seems to us, appropriate and helpful in evaluating the evidence and will be applied when the Tribunal assesses the evidence, both anecdotal and statistical in the following sections of this decision.

V. STANDARD OF PROOF

169. The "standard of proof" determines the degree of probability that must be established by the evidence to entitle the party having the burden of proof to succeed in proving either his/her case or an issue in the case.

170. There are two levels of probability depending on whether the matter to be tried is of a criminal nature, in which case, proof beyond a reasonable doubt is required, or is a civil matter in which case the claimant is required to establish his/her case, or an issue therein, on a balance of probabilities, which is to say a greater likelihood that the conclusion advanced by the claimant "is substantially the most probable of the possible views of the facts." See Duff J. in Clark v. Treking, [1921] 61 Can. S.C.R. 608, at p. 616.

171. The standard applied in Haldimand-Norfolk (29 May 1991), 0001-8 P.E.H.T. by the Tribunal when interpreting Section 5(1) of the Pay Equity Act (Ontario) is contained in paragraph 24 of that decision which reads as follows:

> 24. Having carefully considered the evidence and submissions in this case, we find that the parties have an obligation to ensure the collection of job content information meets the requirements of the Act to accurately identify skill, effort, responsibility and working conditions normally required in the work of both the female job classes in the establishment and the male job classes to be compared. Not only is this a necessary condition of a

> gender neutral comparison system but we also find that section 5 of the Act requires a standard of correctness, that is, the skills, effort, responsibility and working conditions must be accurately and completely recorded and valued. [emphasis added]

172. Section 5(1) of the Ontario Act reads as follows:

> 5(1). For the purposes of this Act the criterion to be applied in determining the value of work shall be a composite of the skill, effort and responsibility normally required in the performance of the work and the conditions under which it is normally performed.

173. Section 5(1) itself does not impose any particular standard which must be met by the parties in order to fulfil the criteria.

174. Accordingly, the decision of the Tribunal in the Haldimand-Norfolk case, insofar as it deals with the standard to be met in the collection of job information is the Tribunal's interpretation of Section 5(1) of the Pay Equity Act (Ontario). It should be pointed out that the issues in that case were whether the employer had adopted a gender biased comparison system and whether it had failed to negotiate in good faith with its employees. The question of the reliability of the results which concerns us here was not directly addressed. That issue relates to the process. The process requires a standard by which to assess the collection of job information and a standard by which to assess the procedures for evaluating that information.

175. The issue before us relates to such matters as the format of the questionnaire, the procedures for gathering information about jobs, the follow-up procedures and safeguards, the composition and functioning of the

committees, the application of the job evaluation plan and the vetting of committee results by statistical analysis.

176. The Commission and the Alliance have advocated a standard of reasonableness to be applied in assessing job information and job evaluation. Also in assessing "damages", by which it is assumed is meant the measure of consequential relief afforded to the complainants by the provision of the Canadian Human Rights Act, a standard of reasonableness is to be applied.

177. Respondent Counsel in his oral submissions when dealing with the onus of proof makes the following statement in Volume 226, at p. 29761, lines 16 - 24:

> Another point on onus of proof is this. The employer's position is that the standard for assessing reliability, the standard for assessing the process to decide whether we have reliability, is one of reasonableness. Did the process work well? It is not a question of whether the process worked perfectly or whether the job information was perfect. The employer has never contended for perfection.

178. In commenting on the Haldimand-Norfolk case, Counsel also makes the following observations in Volume 226, at p. 29761, line 25 to p. 29762, line 13:

> I might just contrast that with the Haldimand-Norfolk case, in which...the Tribunal said, "And we want a standard of correctness." Correctness sounds very like that they were looking for perfection. I don't know...In any event, the employer's position on this is that we are looking for did the process work well. That doesn't mean perfectly.

179. So in the result the parties have themselves advocated a standard of reasonableness. Respondent Counsel's position is that when applying a standard of reasonableness to the results in this study the Tribunal must find that the results fall short of providing a reliable basis on which to render a favourable decision.

180. What standard ought the Tribunal to follow in assessing the reliability of the results? The concept of reasonableness should be viewed in the context of what pay equity or comparable worth hopes to achieve and how it expects to achieve its goal. There are, as well, practical

considerations as to its effects in the work place on the parties involved.

181. Throughout his testimony, Willis, an acknowledged expert in his field, stressed that achievement of pay equity or equal pay for work of equal value as between male dominated jobs and female dominated jobs is not a scientific, mathematical or statistical endeavour. Rather it is an "art" based on a combination of analytical skills, comprehension, intuition and ultimately a subjective evaluation of the job within the framework of the plan while at the same time adhering to the discipline which the plan imposes.

182. In an article by Judy Fudge of Osgoode Law School entitled the "Legal Standard for Gender Neutrality under the Pay Equity Act (Ontario): Achieving the Impossible?", the learned author in referring to a legal standard against which to judge the gender-neutrality of job comparison systems states:

> ...to date there does not exist a conclusive method to demonstrate either gender bias or gender-neutrality in any particular job comparison system. For this reason, the Pay Equity Hearings Tribunal should adopt a reasonableness standard with respect to the issue of gender-neutrality.

183. She then outlines minimum criteria for developing a gender-neutral job evaluation system. It is not necessary to examine those criteria for our purposes since all of the parties to this enquiry have agreed that the Willis Plan initially adopted at the outset of the study satisfies the minimum criteria and is therefore gender-neutral. What is useful for our purposes is to note the observations of the author (at p. 20) where she acknowledges that:

> ...there does not exist a conclusive method to demonstrate either gender bias or gender-neutrality in any particular job comparison system. For this reason, the Pay Equity Hearings Tribunal should adopt a reasonableness standard with respect to the issue of gender neutrality.

184. In commenting on the actual job evaluation process, the author states:

> No matter how scrupulous the design of the job comparison system in avoiding gender bias, bias can creep into the actual process of

assigning job value points to jobs. In other words, the job evaluation system may be fair, but the application can be biased.

185. Fudge then goes on to describe the use of job evaluation committees which if properly constituted following clearly defined procedures would minimize the possibility of bias.

186. Also we refer to the earlier comments of Armstrong that the overall wage gap probably wouldn't disappear completely if pay equity were achieved in all jobs.

187. What is apparent from these comments and from the nature of the subject is that equal pay for work of equal value is a goal to be striven for which cannot be measured precisely and which ought not to be subjected to any absolute standard of correctness. Moreover, gender-neutrality in an absolute sense is probably unattainable in an imperfect world and one should therefore be satisfied with reasonably accurate results based on what is, according to one's good sense, a fair and equitable resolution of any discriminatory differentiation between wages paid to males and wages paid to females for doing work of equal value.

VI. FACTS

A. THE WILLIS PLAN

188. The framework within which work was evaluated during the JUMI Study was through the use of a job evaluation plan.

189. One of the early tasks of the JUMI Committee was to select a job evaluation plan. The JUMI Committee created a sub-committee to examine various job evaluation plans and make recommendations to the JUMI Committee at large. Several plans were examined by the Sub-Committee on a Common Evaluation Plan. In the end, this sub-committee recommended the Willis Plan, designed by Willis, with some minor modifications to better meet the criteria of the Act. Following consultations with representatives of the JUMI Committee, Willis agreed to make changes to the plan including changes to the working conditions chart.

190. The Commission also examined the Willis Plan and expressed its concern about the treatment of "effort" in respect to working conditions. There was also concern about the manner in which the plan dealt with "accountability". Willis agreed to change aspects of the Willis Plan such

that both physical and mental effort would be assessed in working conditions. He also agreed to changes in the treatment of accountability.

191. All participants, including the Commission, appeared satisfied with the changes. Paul Durber, Director of Pay Equity, Canadian Human Rights Commission and a pay equity expert, provided expert evidence as to how the requirements of s. 11 of the Act and the Guidelines were captured by the Willis Plan. Durber stated an essential element of the job evaluation plan, used for purposes of pay equity, is that the tool be gender bias free. In Durber's opinion, there is nothing on the face of the Willis Plan which appears gender biased and there is nothing in the plan that would make it difficult to measure work traditionally performed by men as compared to that of women.

192. The Willis Plan is complex in design. Willis developed this plan in 1974 after working with the consulting firm, Hay & Associates for three years. It uses a matrix format which permits evaluation of the four factors of skill, effort, responsibility and working conditions to be broken down into subfactors. The matrix design allows for one, two or three sub-factors to be assessed on a single guide chart with a total of four guide charts. A guide chart presents the criteria used in the Willis Plan. In some cases, one factor is imbedded in another. For example, interpersonal skills are measured within the levels of managerial skills, thus, how one scores managerial skills affects the number of points given for each of the levels of interpersonal skills.

193. The Willis Plan is a point factor system, which simply means points are assigned to each factor. The point values are added together to arrive at a total point score for each job. The Willis Plan is designed geometrically. Willis has chosen a 15 per cent difference between any two levels of the plan. He finds this percentage of difference is a discernible difference in the semantic definitions of the different levels in the charts. He stated, if the differences were too small, evaluators would be unable to make a choice.

194. The assigning of relative worth to each job is established by the number of points that are available for each factor in the Willis Plan, and there is almost an infinite number of points that are available. The relative number of points that are available for each factor contributes to the conclusion of relative worth of different positions. (Volume 77, p. 9377).

195. Dr. Nan Weiner, President of NJ Weiner Consulting, Inc., a consultant specializing in pay and employment equity, was deemed an expert by the Tribunal in pay equity and compensation. She was asked to express an opinion on the Willis Plan, which she referred to as a system.

Although, she had not worked with the Willis Plan, she stated there was nothing to indicate that it would in any way undervalue female jobs. She indicated a weakness in the Willis Plan could be that, given the breadth and diversity of the Federal Public Service, four levels of interpersonal communication are simply not adequate to differentiate across all the jobs which were evaluated. In her opinion, the Willis Plan attributes more points to knowledge and skill, and accountability or responsibility than to

effort and working conditions, and in some respects, favours white collar work over blue collar work. According to Weiner, it is important for the evaluators who use the system to ensure, through their discussions, the blue collar jobs are measured fairly.

196. In her view, it is not the distinguishing element of the work but how the system is adapted by the user that is important. In this respect, she says in Volume 11, at p. 1564, lines 20 - 23:

> THE WITNESS: What is important is that you ask the people who actually use the system what they did to make sure the system was being used fairly for blue collar jobs.

197. Willis testified the current modified Willis Plan does not have the points on the charts. Thus, the evaluators never know what the points are. They evaluate using pluses and minuses. A computer program determines the points. According to Willis, this frees the evaluator from knowing the point relationships between different jobs.

198. Two elements emerge from the design of a pay equity job evaluation plan. The first is a pay equity plan must be capable of capturing and appropriately valuing female and male work. The second is the allocation of "weight" assigned to the various factors of the plan. For example, in the weighting of factors, the Willis Plan attributes many more points to knowledge and skills than it does to working conditions and physical effort.

199. Willis described his weighting scheme for the Plan. He testified the weights were validated using market rates of pay. Criticism about the use of the market as a measure of validation was expressed by Weiner who stated market influences on the wages for female-dominated jobs are inconsistent with pay equity. Simply expressed, when job evaluation systems undervalue work traditionally performed by women, this becomes compounded in the market place. Accordingly, in her opinion, the market reflects the undervaluation of women's work.

200. The last formal validation of the weights in the Willis Plan was done in 1985. Early in the study, Willis agreed, at the request of the JUMI Committee, to do a validation study of the weights, because of a concern expressed by the Commission. Willis did not believe this re-validation necessary as he had been using the system continuously and had no empirical evidence to demonstrate the factor weights were inappropriate. Management representatives ultimately decided not to perform the validation study because of cost considerations.

201. During the lengthy course of these proceedings, the Employer challenged the weighting of the Willis Plan and its validity as a tool for evaluating jobs in a gender bias free manner. On that basis, the Tribunal heard a considerable volume of evidence. It was not until written submissions from Respondent Counsel were available that the Tribunal and the other parties were advised the Employer agreed the Willis Plan was an appropriate and acceptable evaluation plan for the purposes of the study. Moreover, it was then agreed by the parties that the Willis Plan met the

requirements of s. 11 of the Act and is an appropriate instrument within the meaning of s. 11 for these complaints. Reference is made to the Employer's written submissions at para. 41, p. 11, which read:

> 41. Nevertheless, for purposes of this litigation, the Employer accepts that the Willis Plan was an appropriate plan to use in evaluating jobs in the Federal Public Service. Therefore, the Tribunal need not decide whether weighting the Willis plan is valid.

202. Also, in oral argument, Respondent Counsel stated in Volume 218, at p. 28453, lines 4 - 11 as follows:

> Then, having covered all of those points, we say this: For purposes of this litigation the employer accepts that the Willis Plan was an appropriate plan to use in evaluating jobs in the federal public service. So that, in my submission, was intended to be a complete indication that no issue is raised in respect of the Willis Plan.

203. The Commission has an obligation to assure the Tribunal the Willis Plan meets the requirements of s. 11 of the Act and s. 9 of the Guidelines. In this regard, the Tribunal was assured by Durber with regard to the Commission's view which essentially confirms that the Willis Plan meets the requirements of the Act.

204. During oral argument, all parties agreed on the suitability of the Willis Plan as an appropriate tool for dealing with the complaints before us. Therefore, the Tribunal is persuaded and does find as a matter of fact that the Willis Plan is an appropriate tool within the requirements of the Act and Guidelines for the job evaluations which form the basis for this adjudication.

205. The Willis Plan provides a tool to be used in assessing the relative value of work. But in and of itself it does not give a methodology to determine what is the wage gap between female positions and male positions. The determination of any wage gap is a function of comparing evaluations between male and female jobs. The system itself does not do that without a further step.

B. THE WILLIS PROCESS

206. The Willis Process was developed by Willis over the approximately 24 years he spent as an independent consultant in the area of pay equity job evaluation. The Tribunal heard considerable testimony relating to the implementation of the Willis Process in the JUMI Study. In particular, the evidence covered the period of job evaluation which commenced in the fall of 1987 and concluded in the fall of 1989. In assessing the issue of reliability, we find it appropriate to review each aspect of the process to determine whether it achieved or fell short of achieving the aim of avoiding gender bias.

207. The Willis Process is a process for examining, assessing and evaluating jobs. Participants in this exercise were given the task of measuring the content of each of the positions examined, and asked to assign a value reflective of the total work of each position or job.

208. Although Willis testified the job evaluation plan must be a sound instrument, he also insisted the process within which the evaluation plan is used is more important than the plan itself. According to Willis, everything done in terms of the process was aimed primarily at avoiding evaluations which would suggest traditional relationships or stereotyping. It was designed to avoid anything that might be identified as gender bias. Willis maintained throughout the study, that vigilance during the evaluation stage was of paramount importance, and he continually reinforced the need for objective, fair and equitable evaluations of all positions.

209. By way of historical background, the process eventually agreed upon by the JUMI Committee was not Willis' preferred choice. Willis

initially recommended his proposal for the consideration of the JUMI Committee which outlined processes and procedures to be employed in the study. Due to financial considerations under the control of the management side, his proposal was rejected. Willis then prepared a modified proposal. The JUMI Committee accepted the modified version after a number of "make or buy" decisions were made relating to certain aspects of the Willis Process.

210. The modifications in the data-gathering phase included:

(i) instead of having his consultants conduct the briefing of employees selected to complete questionnaires, he agreed federal employees could be trained to do this task;

(ii) instead of having his consultants review or screen the completed questionnaires, he agreed to train a team of federal employees to perform this task under the direction of a consultant who would be available to oversee this process;

(iii)instead of using consultants to conduct face to face interviews with select incumbents, he agreed to train a team of federal employees to conduct any face to face interviews deemed necessary; and

(iv) in the later stage of the study, Willis reduced the amount of time and involvement that he and his consultants would have in the data-gathering phase of the study.

211. Willis believed these modifications would result in a study which would be of sufficient quality to meet the requirements of the Act based on a number of safeguards he instituted to ensure that complete and accurate job information was obtained.

212. We find it helpful to separately identify each step in the Willis Process, accompanied by the evidence relevant to each step, thereby

assisting us in determining the issue of reliability and the effectiveness of the safeguards.

(i). Data-Gathering

213. Willis testified data-gathering is a critical and most important step in a study of this type. He characterized four possible sources of data-gathering.

214. One source of information is the job description. In reviewing a sample of job descriptions from the Federal Public Service, Willis determined they were out of date and not sufficient for this purpose.

215. A second source is the closed-ended questionnaire, which is described as similar to a multiple choice question. Closed-ended questionnaires require very extensive, in depth and detailed knowledge of the job in order to structure them properly. A great deal of familiarity with the work is required so as to construct the kinds of alternatives provided in a closed-ended questionnaire. This is easier to do in a smaller establishment where there is less variety in the kinds of jobs.

216. The advantage of a closed-ended questionnaire is that it leaves less for the comprehension of the incumbent in terms of their awareness of the range and content of the job. There is instead more reliance on the knowledge and comprehension of the person who structures the questionnaire. The disadvantage of a closed-ended questionnaire is that if the person who structured the questionnaire is not aware of the whole range of the work involved or if they are not fully aware of the kinds of considerations that must go into pay equity questionnaires, then the questionnaires may have fundamental bias built into them. (Volume 180, p. 22971). A closed-ended questionnaire is relatively easy for employees to complete but it is, according to Willis, fundamentally unsound because it permits employees to make value judgments about their work rather than providing factual information.

217. A third source is an open-ended questionnaire which is more difficult to complete than a closed-ended questionnaire. Willis prefers an open-ended questionnaire to a closed-ended questionnaire. Armstrong explained the advantage of open-ended questionnaires is that they are most useful when there is a whole range of quite different jobs to collect information for evaluation. Open-ended questionnaires are also more useful with a literate workforce, which is the case with most of the employees in the public service. (Volume 180, p. 22971).

218. Willis testified he constructed his questionnaire to obtain complete, definitive, accurate and up to date job information. (Volume 68, p. 8542). The data gathered in this study was obtained through an open-ended questionnaire (the "Willis Questionnaire").

219. The fourth and last source of data-gathering involves a task force of professional job analysts who would interview each employee and then prepare the document. Willis has used this approach in a few

instances but stated that it would be "impractical" in the context of the Federal Public Service. (Volume 29, p. 3696).

(ii).Willis Questionnaire

220. Willis discussed the advantages and disadvantages of open-ended and closed-ended questionnaires in the context of a large study such as the JUMI Study. He preferred an open-ended questionnaire as opposed to a closed-ended questionnaire because his aim was to prevent incumbents from making value judgments of their own work. This will occur when a closed-ended questionnaire is employed. Willis states in Volume 65, at p. 8084, lines 12 - 14:

> I think the important thing is that the evaluator must make that value judgment, not letting the employee make it.

221. This questionnaire was used in many previous Willis & Associates pay equity studies in both Canada and the U.S. The JUMI Committee established a sub-committee to finalize the questionnaire's format and content. The amended Willis Questionnaire was agreed to by the JUMI Committee. A guidebook was appended to each questionnaire as a source of assistance to an employee completing a questionnaire. The guidebook was also amended by the JUMI Committee to reflect the Federal Public Service environment.

222. In summarizing his participation in the design of the questionnaire, Willis says in Volume 60, at p. 7429, line 18 to p. 7430, line 6:

> A. The questionnaire has been, I would say, developed over a number of years. It is probably the most worked up questionnaire that is in existence going back all the way to 1974. We have tried to modify it and change it over the years to make it easier for people to complete, but at the same time it is a totally open-ended questionnaire, which I think is necessary.
>
> The final design, of course, was a modification of the questionnaire by a sub-committee from the Joint Union/Management

Committee. I think perhaps we have as good a questionnaire as you could expect to have for any study of this type.

223. Willis participated in the suggested changes in the questionnaires and the guidebook and approved all of the changes which were made. He testified he was satisfied with the questionnaire and the guidebook in the form in which they were used in the JUMI Study. (Volume 62, p. 7654).

224. A portion of the questionnaire provides space for the incumbent's supervisor to make comments. The Questionnaire Sub-Committee had discussed and made changes to this portion of the questionnaire. It was Willis' view these changes were minor and he was satisfied with the questions in their final form. The questions for the supervisor reads as follows:

Carefully review the completed questionnaire, but do not alter or eliminate any portion of the original response. Please answer the questions listed below. We also invite you to consult with your manager on this subject.

1. What do you consider the most important duties of this position and why? (Refer to Question III.)

2. Comment on the accuracy and completeness of the responses by the employee.

3. Please sign on page 34.

IMPORTANT: Significant differences of opinion noted by the immediate supervisor should be reviewed with the employee.

(Exhibit HR-34)

225. Willis stated this kind of check on position information is intended to address two concerns: firstly, the tendency of some employees to overstate their jobs to some degree when there is no supervisor to review the information; and secondly and more importantly, often the supervisor will have additional information which the employee forgets that might be helpful in evaluating the position.

226. One of the problems identified by Willis was obtaining good information from "sophisticated professional level positions." (Volume 68, p. 8544). He stated the higher the level of knowledge and sophistication

of the job, the more it requires the understanding and interpretation of principles and theories, hence greater difficulty is encountered by the individual incumbents in describing their work. In a higher level job it is more difficult for the employee to document and describe their work in a way an evaluator can understand. On the other hand, a very simple cleaning job which follows specific procedures can be documented with relative ease.

227. Another problem encountered in gathering information is ensuring adequate time be given to employees to complete the questionnaire. Each incumbent must be given sufficient time to complete the questionnaire, which is contingent upon the ability of the incumbents to express their jobs in writing. Not only is time an important element in this exercise, but also the effort and care expended by each incumbent.

228. In Willis' expert opinion, the questionnaire was a good tool for obtaining factual up to date job information. In assessing the ability of the Willis Questionnaire to collect sufficient information for the evaluation committees, we note the remarks of Willis in Volume 62, at p. 7686, lines 16 - 22:

> Q. In terms of any of the times that the consultants were sitting in, are you satisfied that by the time those questionnaires came to be evaluated that there was sufficient

> information for those jobs to have been properly evaluated in accordance with the Willis Plan?
>
> A. Yes.

229. The JUMI Committee understood, from the onset of the study, the need to communicate to selected employees the importance of the study as well as the importance of providing thorough job information. As a result, the JUMI Committee established a Communications Sub-Committee to develop a communication strategy emphasizing the necessity of complete and accurate information and a prompt return of the questionnaires distributed for this purpose.

230. The communication strategy included such items as: (i) a pay cheque stuffer explaining the purpose of the JUMI Study and containing an assurance to employees that classification levels would not be affected; (ii) letters to employees who were asked to complete a questionnaire; (iii) preparation of a video for employees designated as screeners/reviewers to

be used in training of incumbents who would be filling out questionnaires; and (iv) training materials for coordinators.

231. Employees were given assurances from the JUMI Committee their participation would not have a negative impact on their careers. They were also assured any information provided would not be used for any other purpose than the JUMI Study. The incumbents from male-dominated occupational groups were instructed that if a wage gap were found, their wages would remain unaffected.

232. To counter possible problems in using an open-ended questionnaire, Willis implements checks and balances, or safeguards, to ensure that the evaluators had complete, definitive, accurate and current information. These safeguards will now be described.

(iii). Coordinators

233. Willis had originally proposed his consultants train incumbents in the completion of the Willis Questionnaire but accepted the JUMI Committee's decision to use coordinators as trainers which he considered a valid "make or buy" decision.

234. The function of the coordinators included training incumbents on how to complete the questionnaires, conducting briefing sessions to explain the nature and intent of the study, responding to employee questions, distributing and explaining the Willis Questionnaire, assisting employees in completing the questionnaires when required, and coordinating the data-gathering process.

235. Coordinators were designated as either national or regional depending upon their purpose and locale. The selection process and criteria applied in selection, used by the Alliance in appointing coordinators was described in detail by an Alliance witness, Elizabeth Millar, Head, Classification & Equal Pay, Collective Bargaining Branch. Similarly, the selection process of the Institute was described by Kathryn

Brookfield, Section Head of Research. On the other hand, no evidence was presented by the Employer as to the manner and criteria for selecting its employees for this role.

236. Coordinator training sessions were conducted in the months of September and October, 1987. Materials, in the form of printed information, slides and videos were given to coordinators to assist in the

training of incumbents. The coordinator training program lasted about a day and a half. Some additional exercises in coordinator training included practice sessions on eliciting the support of individuals who might be reluctant to complete the questionnaire, dealing with language difficulties and making arrangements for interpreters where necessary. All training on the Willis Plan was conducted by a Willis Consultant.

237. With regard to the adequacy of coordinator training, Willis provided the following opinion in Volume 62, at p. 7657, lines 15 - 21:

> Q. In terms of the training, you participated in the training of the coordinators, or your consultants did?
>
> A. Yes.
>
> Q. And you were satisfied with the training that was given to the coordinators?
>
> A. Yes.

238. Following training, each coordinator was then assigned a number of incumbents to train. The date for performing this task was to be decided by the individual coordinator although Willis wanted the training of incumbents undertaken as soon as possible. He also emphasized to the coordinators the importance of having incumbents complete the questionnaires as soon as possible after their incumbent training was given. Willis preferred the questionnaires be completed within a two week period subsequent to the employee training. Willis estimated it would take incumbents four to eight hours to properly complete the questionnaire.

239. Following the training of the coordinators which was completed in October of 1987, approximately two-thirds of the questionnaires were received by February of 1988 and up to three-quarters were received by March of 1988. The Administrative Sub-Committee, established by the JUMI Committee, spent considerable time assessing the number of questionnaires which had been received and ways and means of obtaining all the remaining questionnaires. The final rate of return for the questionnaires was 95 per cent. A few questionnaires continued to come in over the summer and fall of 1988.

240. The Tribunal heard from Brookfield who testified that many of the coordinators from the Institute commenced their training task very soon after receiving coordinator training. She explained that the employee training went on for a considerable period of time because the coordinators had a large number of employees to train and the employees were not all at a single work site. These factors required staggered training sessions for

coordinators to meet with different employee groups. Brookfield also indicated some of the incumbents could not be released from their work at the same time, and this factor also lengthened the period of time required for the training.

241. Brookfield also expressed the Institute's view as to the calibre of training provided at the end of coordinator training. She said in Volume 168, at p. 21007, lines 5 - 16:

> A. They said to me quite frankly the more they did it, they felt the better they got and that they had received input from previous training sessions about questions that employees would have and they would respond to them at that point. But then, after that, they might think of more information or another way they might have addressed that concern and they would incorporate it in their next training session, perhaps up front, or be able to raise, if there weren't questions, possibilities and things they had gleaned from other training sessions.

242. The Alliance had many more coordinators than the Institute, numbering approximately 100. Margaret Jaekl, Classification and Equal Pay Officer, Collective Bargaining Branch, of the Alliance, testified as to the effectiveness of coordinator training from feedback she received from Alliance coordinators. Jaekl states in Volume 200, at p. 25831, line 25 to p. 25833, line 8:

> Q. Did you receive feedback from the co-ordinators as to how they felt their role was being received, first of all, by management and, second of all, by those that they were training in the filling in of the plan?
>
> A. Yes. We had meetings from time to time with all of what we called our national co-ordinators. Each component had a national co-ordinator and then they had many regional co-ordinators, too.
>
> ...
>
> A. The feedback we got generally was that they felt they were working well with their management counterpart. People were understanding their presentations. People were generally completing their questionnaires and returning them. Some people had questions and, in general, they felt comfortable that they were able to answer those questions.

243. The JUMI Committee sought cooperation from management in granting time to employees for training. The uncontradicted evidence is there was good cooperation from the Employer in providing the selected employees with sufficient time during normal working hours to attend the training session and to complete the questionnaire. Incumbents were given time off with pay to complete the questionnaire which could involve up to eight hours, where necessary.

244. In Willis' opinion, the shorter the time lapse, between the training of coordinators and their training of the incumbents, the more effective the training would be. Willis' experience was he was able to track the quality of the questionnaires in terms of how soon the incumbents completed the questionnaire after receiving their training from the coordinators. According to Willis, quality goes downhill over time. In this particular case, Willis was not able to pinpoint when the quality began to decline. He found a variety of quality levels in the completed questionnaires. He remarked the earlier questionnaires possessed a higher quality. The Department of National Defence questionnaires were completed right on schedule. Willis testified these employees completed the questionnaires as they were supposed to be done and were "excellent questionnaires". Willis and his consultants noticed a "dropping off" in quality the longer it took for the questionnaires to be returned.

245. There is little evidence concerning specific dates of coordinator-incumbent training sessions. Some portion of the delay can be attributed to the time supervisors took to read, comment on and sign employee questionnaires. Some supervisors waited until all of their employees had completed their questionnaires and signed them off en bloc. Willis admitted there was no way of knowing whether an employee had, in fact, filled the questionnaire out within the goal of 10 to 14 days after receiving their training or at a later time. It is noted there were 1,258 incumbent substitutions in total involving 837 questionnaires.

246. The evidence revealed the information from female employees came in sooner and was of better quality than the information received from male employees. Also, questionnaires from incumbents of high level technical and professional positions were returned later and contained weaker information than questionnaires from the incumbents of clerical and vocational positions.

(iv).Screeners and/or Reviewers

247. As the completed questionnaires were returned, one of the Willis Consultants, Jan Drury, was asked by Willis to select the best questionnaires for evaluation by the MEC. Drury expressed concerns to Willis about the overall quality of the questionnaires. As a result, Willis then instituted a back up procedure to obtain additional information. This involved a task force of employees, appointed by the JUMI Committee, referred to as screeners and/or reviewers. Their primary function was to screen incoming questionnaires for any gaps in information and/or inconsistencies.

248. According to Willis, the screening of questionnaires is an absolute necessity in the Willis Process. It was Willis' original recommendation for the study that the consultants perform the screening and reviewing function. Normally, Willis would use his consultants to screen the completed questionnaires. The JUMI Committee decided to train federal government employees to perform this task. This triggered another "make or buy" decision by the JUMI Committee. The screeners/reviewers functioned throughout the duration of the study.

249. The screeners and reviewers were trained by Drury. They received more extensive training than the coordinators because the screeners and reviewers had to be familiar with the Willis Plan in order to assess whether the questionnaires were properly completed.

250. Accordingly, the management side and the union side each appointed individuals to act as screeners/reviewers. Approximately 55 individuals functioned in this capacity. Their responsibilities included undertaking certain technical tasks for each questionnaire, such as removing all gender and classification references. After identifying questionnaires requiring additional information or clarification, the screener/reviewer was then required to draft questions to ask incumbents in order to complete the necessary information. They also obtained further factual information respecting technical terminology found in the questionnaires and presented this information in terms better understood by an evaluation committee.

251. Drury oversaw the work of the reviewers until March of 1988. Drury examined the review questions and notes drafted by the screeners/reviewers for each review completed on the questionnaires evaluated by the MEC. Subsequently, Diane Saxberg, on the union side, and Doug Edwards, on the management side were appointed Chief Reviewers as of March 7, 1988. The Chief reviewers were responsible for reviewing the draft questions of the screeners/reviewers.

252. The screeners/reviewers interviewed the incumbents to obtain the required information. A high percentage of these follow up interviews were done by telephone and only a limited number, less than a dozen, were done in person. In some instances, obtaining this information required several telephone calls, some of which were extremely lengthy. The responses were then written up and appended to the questionnaires before being conveyed to an evaluation committee. The written responses were referred to as "reviewer notes".

253. Willis wanted the screeners/reviewers to identify areas in the questionnaires where something may have been overlooked or left out, or where there might have been contradictions between what the incumbent wrote and the comments of their supervisor. They were also instructed to be alert to expressions of opinion or conclusion not supported by fact.
254. The screeners/reviewers found only a "handful" of cases in which there was disagreement between the supervisor and incumbent. Saxberg testified in these situations she would talk to both individuals and, in most cases, reported the disagreement was more of a semantic nature than a substantive disagreement about job duties.

255. Willis explained, based on his past experience, about 50 per cent of the necessary interviews can be conducted by telephone but the other 50 per cent require a personal meeting, in order to obtain more substantive information, especially when dealing with higher level technical and professional jobs.

256. Willis stated the number of times that a questionnaire has to be supplemented, whether it is 80 per cent or 30 per cent of the cases, does

not really impact on the quality of the questionnaire. According to Willis, it is the extra information which is obtained and put before the committee that counts.

257. The evidence indicates there were some reviewers who previously had functioned as evaluators on an evaluation committee and who had been identified as "outliers" in terms of their evaluations. Willis defined an outlier as an individual, on an evaluation committee, who exhibits a divergence from the rest of the committee as a whole and gives higher scores to certain kinds of jobs or lower scores to certain kinds of jobs compared to the other members of the evaluation committee. (Volume 29, p. 3793).

258. Willis indicated one way of checking for validity, in the situation where a screener/reviewer is also an outlier, is to examine the questions they draft and the answers they give and determine whether the answer responds to the question. Willis testified he saw no indication these individuals were not recording the answers to the questions raised.

259. The Tribunal heard evidence from three individuals who performed as screeners/reviewers. With regard to the effectiveness of telephone interviews for obtaining information, Christine Netherton states in Volume 173, at p. 21919, lines 5 - 20:

> A. ...sometimes it only took half as long to get the information, but very often you would have to explain what the study was doing and they would say "Oh, I filled that out" and so and so. So there would be a lot of chat to get easy with them. And you tried not to rush people.
>
> I think the information did come back on the whole. And you would get this response from other reviewers as well.
>
> But there would be the person that did not like talking. I am talking of the impression I am left with. I am not saying that it was 100 per cent perfect. But the main impression is that in the majority of cases you did get good information via the telephone.

260. Another reviewer/screener who testified, Mary Crich, said she did not often see examples of conflict in the summary of duties and responsibilities between the incumbents and the supervisors. Crich gleaned from the telephone interviews that employees enjoyed the opportunity to speak with someone about their job.

261. Both Willis and Durber were asked about the competency and ability of the screeners/reviewers. A number of individuals who functioned as screeners/reviewers were familiar to Durber because of his lengthy experience in the Federal Public Service. Durber described them as "professional job evaluators as well as analysts". In his opinion, they would tend to be more competent to perform the tasks assigned to them as reviewers than others without similar backgrounds. (Volume 164, at pp. 20505-07).

262. With respect to the adequacy of their work, Willis said the following about screeners/reviewers in Volume 65, at p. 8136, lines 2 - 7:

Q. But were you aware that after training reviewers had any difficulty understanding their job?

A. I don't believe any of them had any difficulties. At least none were expressed to me.

263. Following the screening/reviewing process, the questionnaires with the reviewers' notes would be turned over to an evaluation committee. If an evaluator on the committee required further information, questions would be drafted by the evaluation committee and would be passed back to the screener/reviewer to solicit the necessary information from the incumbent. The information obtained by the screener/reviewer would then be provided in writing and returned to the appropriate evaluation committee.

264. Under the direction of Durber, the Commission examined questionnaires with a view to assessing their quality. During the hearing, the Commission introduced a report, An Examination of the Quality of Questionnaire Information used by the Federal Equal Pay Study (Exhibit HR-245). The report which examined the quality of questionnaire information was prepared by the Pay Equity Directorate of the Commission at the request of Durber, the investigator into these complaints. An experienced researcher, who possesses a Master's Degree in Canadian Studies from Carleton University, was commissioned to review a cross section of the evaluations. This included 63 benchmark evaluations and 588 non-benchmark questionnaires, a total of 651 questionnaires. Her task was to ascertain the apparent completeness and accuracy of all material in the questionnaires files collected as part of the JUMI Study. The researcher was closely supervised by Durber. As part of this work, Durber personally reviewed 36 files which were flagged by the researcher and found each to be in satisfactory condition.

265. The researcher reported the legibility of the descriptions in the questionnaires was good in all cases and that the open nature of the questionnaire appeared to provide scope for answers for both male- and female-dominated occupational groups. Many incumbents enlarged on their duties by adding pages to this portion of the questionnaire.

266. The Commission's report also recorded supervisor signatures were affixed to over 99 per cent of the questionnaires and in over 96 per cent of them the supervisors provided comments. Contradictory information from supervisors appeared in approximately 9 per cent of the questionnaires. In the questionnaires where supervisors provided conflicting information, 95 per cent were resolved by subsequent interviews conducted by the screeners/reviewers.

267. Durber expressed his own expectations about the quality of the questionnaire information when he said in Volume 158, at p. 19761, line 23 to p. 19762, line 3:

> I can only say that from my experience in the public service, what I did see was much superior to what I have seen in job descriptions and in job files, presentations even in grievance situations, just to try to put my own expectations into some sort of context.

268. During cross-examination by Respondent Counsel, Willis was asked whether the safeguards implemented by the JUMI Committee to address problems in the data-gathering stage achieved what he wanted. Willis said in Volume 78, at p. 9543, line 3 to p. 9546, line 1:

> Q. Those safeguards -- and they were all described in your original proposals -- related to both information-gathering and evaluation. Right?
>
> A. Yes.
>
> Q. I am going to suggest to you that almost or wholly without exception the safeguards that were implemented -- and there were lots -- were not effective to achieve what you wanted them to do.
>
> A. I think it is fair to say that there were degrees of effectiveness that I experienced.
>
> Q. And the degree of effectiveness, I am going to suggest to you, is disappointing at best.
>
> A. Yes.
>
> Q. Part of the result of that is that when we come to the information that was made available to the five and nine committees, after all the shoring-up it was weaker than it should have been. Do you agree?
>
> A. I am not sure what you mean by "weaker than it should have been".
>
> Q. Weaker than is desirable for a good evaluation.

A. I did feel -- and I expressed this to the Joint Union/Management Committee -- that the quality of the information was not as high as I would have liked. However, I felt that overall it was satisfactory for our purposes.

Q. I understand, but you have also told us that it was weaker than what you normally get in other studies.

A. Yes.

Q. Even because of some of the weaknesses in the safeguards we have to raise something of a question mark or a flag, if you will, over some of the information that was actually obtained, some that is actually there, because of some of our discussion that it

wasn't written by a skilled job evaluator and some of the entries were made by outliers; all that discussion that we had. Do you agree with me?

A. Are you suggesting that some of the information may have been inaccurate?

Q. I am not saying that it is inaccurate. We don't know whether it is accurate or not. Our level of confidence in the information is below what we would like because the information is, to some extent, written by people who aren't skilled in doing this kind of writing, it was screened by people who aren't skilled in screening, interviews were conducted by people who aren't professional job analysts. That is what I am saying.

A. I think I did express to the Joint Union/Management Committee, or at least to the Mini-JUMI, that we would expect a wider amount of disparity because of the information being somewhat weak.

Q. But what I am suggesting to you, in addition, is that -- you say "weak". I am asking you whether you agree with me that even with what we have, we have a somewhat reduced level of confidence in its accuracy.

A. I don't know if I can say that. Certainly what we would have liked would have been questionnaires that were more complete and that focused more on factual information. These are the

things that always lead the evaluators to making certain assumptions, resulting in a wider range of possible disparities.

269. In spite of the fact the quality of information was weaker than what was available to him in other studies in which he had been involved, Willis consistently maintained throughout the course of this hearing that the quality of the information was good enough for the purposes of this study.

270. We note, in the course of further cross-examination by Respondent Counsel, Willis again gave his opinion on the quality of the information. This response is found in Volume 69, at p. 8612, line 22 to p. 8615, line 15, where he says:

Q. So what you are saying is that after all the shoring up, the information was still wanting to some significant degree.

A. I would say that the information at best was satisfactory, but not superior.

Q. Would there have been a range -- you make me think of our performance appraisals. We can get "satisfactory", "fully satisfactory", and "superior". Is that the table you are using?

A. Let me put it this way: I had some concerns about the quality. Telephone interviews and interviews by interviewers who were not professionally trained can never completely substitute for a well-completed questionnaire in the first place. While overall I would say the quality was sufficient for our purposes, particularly with the large numbers of evaluations -- again, if this had been Addiction Services with only 19 or 20 positions, I would have been very concerned because I knew that we had to tolerate a greater disparity than I would have hoped for as a consultant. Again, as long as the disparity is random and it will cancel itself out in the end, I felt that I could live with the result.

What happens when you do have a questionnaire -- two things happen when you have a questionnaire that is somewhat weak: (1) it slows down the process, as we found out; and (2) we have to anticipate that there will be a wider tolerance for disparity.

Q. All right. I want to stop you there because you said something again that I want to challenge you on.

You are saying that the disparity cancels itself out. That is if you are looking for gender bias.

A. If it is random, by definition it will cancel itself out. If there is a pattern that results, then it isn't random.

Q. I am going to suggest to you that what it does -- if it is random, it cancels out gender bias.

A. It will cancel out any bias.

Q. Any bias, all right. But what it doesn't cancel out is unreliability. If you have extensive disparity, what you have is a lower level of reliability. I thought we agreed on that yesterday.

A. I think a statistician would say that if you were dealing with a relatively small number, that would be very true. It is less true as the number of evaluations grows, and the disparity continues to be random -- that is, the pluses and the minuses tend to cancel each other out -- you can still achieve satisfactory reliability with a large number of evaluations.

271. We also note, during cross-examination by Respondent Counsel, Willis reiterates his previous testimony in Volume 78, at p. 9566, lines 19 - 22:

I think I have already said that I feel that with all of the work we did on the data gathering that the data is good enough for the purposes of this study.

272. And again at p. 9567, line 23 to p. 9568, line 14:

Now I am saying to you, Mr. Willis, what do we have to take away before you would say, "I will not defend this study"?

A. Number one, I did make the statement several times that the quality of information was good enough. I would have blown a whistle if I felt that the quality was so low that we couldn't depend on it.

Second, I did indicate that I felt strongly that I could not validate the results of the study if we couldn't do an assessment, an internal review of existing evaluations. That has been done.

But what it would take? It is possible that I might look at that final analysis and say that I agree that we cannot use the results, but I don't know that.

C. THE EVALUATION PROCESS

273. In a large study such as the JUMI, involving a significant number of positions, Willis utilizes multiple evaluation committees. One committee is his preferred approach, but with myriad jobs, it is necessary to rely on more than one committee in order to evaluate efficiently and properly. Overall, there were 16 evaluation committees established to evaluate questionnaires.

(i). Master Evaluation Committee

274. The challenge for Willis is to design a process which enables the various committees to be consistent with one another over a relatively long period of time. As a guide and procedural safeguard, Willis creates a steering committee or a master evaluation committee. Willis stated it is necessary and essential in a pay equity exercise to make comparisons among dissimilar jobs. The master evaluation committee has the primary responsibility for establishing the relationships among different jobs and setting the frame of reference for the multiple evaluation committees. This exercise is what may be described as the master evaluation committee "discipline".

275. The MEC evaluations are referred to as benchmark evaluations. The MEC evaluated a total of 501 evaluations. Benchmark evaluations are critical in a process where multiple committees are used.

276. The MEC was composed of 10 members, one half management representatives, and the other half union representatives. One management representative and one union representative were designated as co-chairs. Willis did not select the MEC members, this was left to the parties discretion. Willis recommended its members have a government wide perspective of work performed, analytical/conceptual skills, dedication to completing a tough assignment and an ability to submerge feelings of union or management affiliation in order to achieve a balanced approach to evaluations. The parties attempted to structure the MEC to reflect that balance.

277. Willis testified the MEC had a good balance of males and females with a good variety of backgrounds. The MEC also had an even number of union and management representatives.

278. According to Willis, the key to successful job evaluation is consistency in the interpretation of the evaluation factors as between the multiple evaluation committees. He uses three methods to test for consistency, all of which were employed in the JUMI Study. The first method is used in situations where a consultant is facilitating an evaluation committee or acting as an advisor. Here the consultant independently evaluates the same job using the same information the committee members are absorbing, while at the same time looking for committee patterns which may differ from the independent consultant evaluation. The second method consists in comparing individual evaluators to the committee as a whole. Finally, the third method consists of comparing committees to one another. Testing for reliability between evaluators, inter-rater reliability, and testing for reliability between committees, inter-committee reliability, will be described and examined in greater detail in a following section.

279. Benchmark evaluations provide a broad frame of reference for evaluation committees and are utilized to achieve consistency and function as a kind of quality control in the evaluation process. More specifically, the term "discipline" refers to the liberalness or conservativeness with which the MEC interprets the evaluation semantics.

280. The consultants must ensure the discipline is consistent, among the different evaluation committees. The discipline adopted by the MEC places a heavy responsibility on the multiple evaluation committees to evaluate the jobs and ensure they track well and are consistent with the jobs the MEC evaluated. That is, if the MEC evaluates a certain factor in a certain way, it must be adhered to by the other evaluation committees.

281. Willis testified if the multiple evaluation committees were permitted to create their own discipline, the end result would be that the evaluations would be inconsistent. The evaluations might be consistent within themselves, that is, the multiple evaluation committees might treat all jobs fairly and equitably, but the degree of liberalness with which they interpret the semantics might differ. If the master evaluation committee evaluates a factor in a certain way, that same approach must be adhered to by the multiple evaluation committees, otherwise over or under evaluation of questionnaires arises. In Volume 60, at p. 7396, line 18 to p. 7397, line 4, Willis stated:

> Every evaluation committee adopts what I have referred to as a discipline, which is a conservativeness or liberalness in treatment of the evaluation factors. Once that discipline is established, if an evaluation comes in higher or the job is evaluated more liberally than the discipline would suggest by other evaluations, I would call that an over-evaluation. If the evaluation was more conservative than I would have expected compared with the overall consistency of the committee, then I would call that an under-evaluation.

282. Willis felt it critically important the MEC provide a sound evaluation basis for the other committees to use as a frame of reference. As to the relative quality of the questionnaires used by the MEC as compared to those used by the other committees, Willis stated that the quality of questionnaires used by the MEC was higher than used by the other committees.

283. Willis requested Drury select benchmarks for the MEC, based on a broad representation of the depth and breadth of the organization. The JUMI Committee formally approved the criteria for selection at its July 10, 1987 meeting. These criteria specify that benchmark positions would be representative of all occupational groups, different organizational levels, high population jobs, standard jobs, and mix of male- and female-dominated occupational groups in the total study population sample. As well, care would be taken to ensure that there was a sampling of specialized positions and that consecutive levels within a job series would be minimized.

284. He also gave Drury another criterion for selecting benchmarks which was to pick questionnaires of the highest quality. Quality in this context, according to Willis, was completeness, definitiveness and factual content. Willis felt it was very important the MEC have the highest quality questionnaires.

285. The MEC did not enjoy the luxury of receiving all the questionnaires beforehand and then selecting those to be used as benchmarks. Willis was instructed by the JUMI Committee to begin the MEC's work as soon as the first 50 questionnaires were returned. In fact, some questionnaires were still being returned when the MEC had finished its work. While Willis was satisfied overall that the MEC provided a good frame of reference, he could not say each of the criteria approved by the JUMI Committee for selecting the MEC's questionnaires was satisfied in selecting the benchmarks.

286. At the beginning of the MEC's work, Willis functioned as the chair of the committee. After a period of time, Willis relinquished the role of chair to the MEC co-chairs who rotated on a weekly basis. The role of the chair was to facilitate the meeting, maintain a neutral posture so as not to influence the group, write the evaluations on the blackboard and lead the group through the consensus process. Willis spent some time with the co-chairs, coaching them as to what he was doing, and why he was doing certain things. It was about three weeks before they assumed this task. From that point on Willis sat in the back of the room as an observer and was called upon from time to time for interpretation. He also functioned as a facilitator during the "sore-thumb" or "interim review sessions" (another part of the process which will be explained later). He proceeded on that basis all the way through. Whenever it was time for a review session, he would take over from the group.

287. After the MEC had completed its work, Willis suggested, for efficiency purposes, that a portion of the MEC benchmarks be designated "primary benchmarks". As the additional job evaluation committees began their work, they required access to the MEC benchmarks. Rather than having a complete set of all benchmark evaluations available to each evaluator,

primary benchmarks were identified and provided to each individual evaluator. However, each evaluation committee was provided with one complete set of benchmarks.

288. The selection of primary benchmarks was based mainly on expected frequency of use and on other factors such as different organizational levels, different occupational groups, and the inclusion of different factors which were most representative of the jobs evaluated. At Willis' request, each of the MEC members produced a list of benchmarks, which was refined by Willis and in the end approximately 100 primary benchmarks were identified.

(ii).Multiple Evaluation Committees

289. Each of the remaining multiple evaluation committees had seven members equally divided between union and management. One member functioned as either a management or union chair of the committee. Again, Willis left the selection of these members to the parties. The Tribunal heard evidence from the Alliance and the Institute that care was taken to select individuals who were articulate, analytical, able to defend the evaluations and willing to work as a team. In terms of balance between the sexes, the Alliance attempted, without success, to recruit equal

numbers of males and females. Their female evaluators were, however, often members of male-dominated occupational groups.

290. Willis believes a mix of genders on a committee is important primarily because of perception. As he said, if a committee is all female, it could be viewed as a female-oriented study or might be perceived the other way if the committee was all male. Willis' experience is if a committee has "good" people on it, their gender is not important. Willis considers the background of the members more important than the sex of the individual doing the evaluation.

291. Willis had recommended that no Federal Public Service classification specialists be on evaluation committees. This, however, did happen. Seven evaluators nominated by the Employer had extensive knowledge of the classification system in the federal government. They served on four evaluation committees and on the MEC. Willis' concern about individuals with classification background is they tend to bring what he refers to as "baggage" to the evaluations. Willis believes someone who is totally inexperienced will likely be more objective than someone with years of experience in classification.

292. In this context, Willis described "baggage" as pre-existing knowledge and understanding of the relativities within an organization. For example, baggage refers to assumptions about work and are probably unconscious. He views baggage as biases based on incomplete information from which hidden agendas could arise because of those beliefs.

293. Everyone carries "baggage" of one sort or another, according to Willis. It can, nevertheless, be minimized with an open mind and an objective, fair attitude when applied equally to all jobs so as not to improperly influence an evaluation.

294. Each of the five and nine evaluation committees consisted of seven members at all times; however, in many instances, substitutions did occur. The Tribunal heard direct evidence from 17 evaluators.

295. There was testimony from one of the 17 evaluators, Christine Netherton, a member of the first version of Committee #1 (it functioned after the MEC was finished) concerning the element of baggage. One member of her committee had a classification background. Netherton testified this particular individual had difficulty appreciating other points of view because of her background in classification. When this kind of problem emerged, the committee would attempt to discuss it with the member.

Failing a resolution of the problem they would obtain the assistance of a consultant.

296. This problem was also identified by Willis with evaluators on Committee #3. The first formation of Committee #3 had numerous problems. Some of these can be attributed to the fact that certain of the management evaluators had former classification backgrounds. On the staff side there were evaluators committed to raising the scores of female-dominated occupational groups higher than was warranted. Willis described the way this committee functioned as "almost a stand off". Further details of the problems in Committee #3 are canvassed in Willis' evidence in Volume 57, at p. 7090, line 12 to p. 7093, line 20:

> A. Number 3 had some individuals on it who, on the staff side, were people who seemed to be committed to having the jobs of people in female occupational groups up as high as they could and two of the three on the management side were former classification people and they seemed to be devoted to keeping them as much in line as they could. It was almost a standoff.
>
> The Chair of Committee #3 was a union representative and, while we counsel the chairs very carefully to take a neutral position -- that is, the chair, for their own credibility and not to have an undue influence, should be very careful how they led or how they facilitated the groups -- this particular chair almost became a fourth union evaluator. Not that she actually evaluated, but she entered into discussions in a way that leaned toward the union side rather than taking a neutral posture.
>
> Of course, the chair has the opportunity of consensing and moving on. She would never move on until her side seemed to be well represented. This was one case where I felt that it was imperative that the chairperson be removed and I so recommended to the Joint Union/Management Committee.
>
> Q. And what happened with regard to your recommendation?
>
> A. Nothing. The management side supported my recommendation, but the staff side refused to go along with it.
>
> Q. So, what was the result of this standoff? How do you feel it affected the evaluation process within Committee #3?

A. Interestingly enough, they tend to be a good match. One of the management side was a former classification manager and he was very forceful. It turned into a standoff in most cases.

The problem was that they were evaluating slow and slower. While I would expect eight or nine evaluations a day, I was not able to get that sort of productivity from any of the committees. But this particular committee was evaluating two or three jobs a day and they were themselves becoming extremely frustrated. So, I felt that the exercise was detrimental not only to productivity but to the health and well-being of the members themselves.

Q. You have mentioned two consequences of this standoff, being the health of the committee members and also the slow productivity rate. What is your opinion with regard to the actual evaluations performed by that committee?

A. I can't say that we found any pattern of bias that grew from that committee. I am sure that we would have gotten a pattern if it hadn't been three on one side and three on the other. I think the evaluations, at least as far as we could determine, were okay.

Q. What eventually happened to Committee #3?

A. At the time that we expanded from five to nine committee [sic], we were able to remove the chair from the leadership role and place here as a voting member of one of the other committees, one of the nine committees. Surprisingly enough, her attitude seemed to improve dramatically at that point.

Q. What do you mean her attitude improved?

A. In the opinion of the consultant sitting with this committee and in the opinion of the chair of the committee, she was handling herself more conscientiously than she had.

Q. So, with regard to the evaluations performed within the subsequent committee she worked on, do you have an opinion to give on that?

A. I didn't see a pattern of problem developing either with the individual or with the committee itself.

297. The best pay equity evaluation results are obtained from having truly heterogenous committees. The ideal profile of a job evaluation committee in this kind of exercise is to have individuals with different

backgrounds, different experiences, with approximate numbers of males and females, individuals representing different unions and employees with different functions representing different departments and organizational levels. Willis' goal is to obtain individuals who can be called upon to evaluate conscientiously and fairly, not an easy task.

298. According to Willis, bias can occur in the use of job evaluation systems, not necessarily from the evaluation plan but on the part of an evaluator. This is the reason he thinks that the process itself is more important than the job evaluation instrument. A heterogenous committee cannot guarantee bias will not creep into a job evaluation process; however, with this kind of committee, there is a better chance of getting an objective result. As Willis states in Volume 29, at p. 3788, lines 18 - 22:

> ...we need people who are conscientious, who can be analytical, and who could be depended on to do their best to do what is right, rather than to protect their own particular field or area.

299. The Tribunal heard evidence on the backgrounds, ages, positions held, skills, strengths and weaknesses of the members of the multiple evaluation committees. With some exceptions, the evidence generally supports Willis' criteria of a balanced committee. On the other hand, the evidence clearly indicates there were some evaluators who carried baggage, who had agendas, and could not be depended upon to evaluate jobs objectively. To the extent these evaluators may have affected the reliability of the results, we will review other procedural safeguards used by Willis in the evaluation process to determine how well the process worked.

(iii). Process for Evaluation of Questionnaires

300. Willis described the difference between what is commonly known as traditional job evaluation and pay equity job evaluation. Willis stated traditional job evaluation has been used since the early 1940s to evaluate primarily management jobs. Its purpose is to achieve some basis for applying pay differences among different levels of managers. On the other hand, Willis states pay equity requires comparisons of dissimilar jobs at all levels within an organization and in the market.

301. The methodology tends to vary considerably between traditional job evaluation and pay equity job evaluation, although they both utilize evaluation committees. Routinely, in traditional job evaluation,

committees are made up of managers, information is collected using job descriptions and interviews are conducted by consultants. The consultant's role becomes less intrusive once a committee is trained to evaluate. As Willis says, "the learning curve goes up rather dramatically."

302. In pay equity job evaluation, Willis prefers committees that are: (i) "balanced", comprising equal representation of male and female; (ii) with cross-sections of different organizational levels; and (iii) representative of diverse backgrounds.

303. Willis further testified pay equity job evaluation committees have to be trained in how to look at a questionnaire and to analyze the importance of a job. During this process, they must submerse their personal feelings about how jobs tend to fit together. Willis stated different problems are encountered in pay equity job evaluation than in traditional job evaluation. Primarily, problems arise in pay equity because of "people's feelings about job relationships."

304. Willis finds evaluators are comfortable in the context of traditional job evaluation because of their general understanding of jobs, as for example, a group of managers evaluating jobs in their own organization. Thus, "peoples feelings about job relationships" become less important in that context than in a pay equity job evaluation process where the consultants have to actually get the evaluators to look at things differently than they ordinarily would.

305. The Willis Process requires evaluators to evaluate independently. The Willis Process prescribes a particular procedure which must be followed during evaluations. The procedure may be described as follows: each member of the committee reads the questionnaire on their own; then the evaluators discuss the information and raise questions about job content which Willis equates to the final step in the data-gathering process; during the discussion stage, Willis permits committee members to share any special factual knowledge about the nature of the work performed and the context in which it is performed; should any evaluator or committee require additional information, questions would be drafted at this time and sent to a reviewer; when the committee members have a common understanding of the facts, each evaluator is required to independently and confidentially, rate each factor pertaining to that position; subsequently, the consultant or chair collects all of the evaluation slips which contain each individual evaluators rating and transfers them to a blackboard, thus giving the committee a visual basis for making comparisons.

306. There follows a discussion period in which the evaluators talk about their evaluation differences; if an individual has a slightly different rating for any given factor they are called upon to justify, in factual terms, their rating; Willis expects the other members of the evaluation committee to listen to the reasons of minority evaluators and he refers to this part of the process as the "consensus process"; he then permits individual evaluators to adjust the factors at this point, but only if they can demonstrate factual reasons for this adjustment; the consensus score is recorded and a rationale is prepared which explains essentially the reasons for the particular evaluation of each position, using criteria defined in the evaluation plan, and exemplified by the benchmarks.

307. Although Willis advised the JUMI Committee it was only necessary for the MEC to prepare rationales, the JUMI Committee decided that the multiple evaluation committees should prepare written rationales as well. Problems arose in relation to the rationales as some were poorly written and difficult to decipher. Consequently, there were delays in their transcription. In the past, Willis has not used rationales because he does not consider them critical to the evaluation process although, he did testify they can be helpful. Willis counselled multiple evaluation committee members to use the rationales as a guide but, in every case, he wanted the evaluators to return to the MEC questionnaire and read it, rather than relying on a rationale. It was Willis' opinion it was impossible to capture all the things that an evaluator would need to know in order to evaluate a position in a one or two page rationale.

308. Either before reaching consensus or after reaching consensus, depending upon the preference of the individual committee, the committee

looked at the MEC benchmarks and selected either similar or dissimilar jobs to ensure their ratings were consistent with the MEC benchmarks. If their evaluation scores were inconsistent with the MEC benchmark then the committee had to adjust its evaluation to accord with the benchmark evaluations.

309. Willis stated it becomes fairly obvious, particularly to the consultants, when an individual demonstrates a gender preference during this process because it is difficult for an individual to provide factual information to support a preference based on feelings. Willis does not require unanimity for consensus, but requires a two-thirds agreement by members of the evaluation committee. Any evaluator in the one-third minority has an opportunity to persuade the group their rating is the correct one. As Willis stated, "...in the final consensus, we have to have

at least two thirds of the people who feel the evaluation is right." (Volume 38, p. 4737).

310. The evaluation committees tended to follow the evaluation process designed by Willis, that is, independent reading of the questionnaires, discussion among committee members to better understand the facts, individual rating of each subfactor, posting of individual ratings on a blackboard, arriving at a consensus and selecting appropriate MEC benchmarks.

311. Willis testified a good committee possessing good job information can usually evaluate 8 to 10 questionnaires per day. On that basis, the JUMI Committee initially established 5 multiple evaluation committees, with the expectation that each committee would be able to evaluate approximately 750 positions; however, productivity was much lower than originally anticipated for the MEC and the multiple evaluation committees. Consequently, in order to deal with the time delay and to solve other problems, Willis recommended and the JUMI Committee agreed on March 3, 1989 to reform and expand the five multiple evaluation committees to nine.

312. Many of the problems observed by Willis and his consultants occurred with the initial five evaluation committees. The circumstances surrounding these concerns are now detailed.

(iv).Training of the Multiple Evaluation Committees

313. The evaluators needed training in the use of the Willis Plan. Willis personally trained the MEC in October, 1987. Willis testified he was satisfied with the training of the MEC. (Volume 62, p. 7698).

314. Training of the first five evaluation committees was undertaken by Willis and his consultants. Willis met with all five evaluation committees for the first day, and thereafter he divided the members into evaluation committees and assigned a consultant to each committee. When the five evaluation committees expanded into nine, all new committee members received individual training or, if it was a new fully constituted committee, then training was undertaken with the whole committee.

315. One of Willis' goals in training a committee is to ensure comfort with the Willis Plan. His training usually consists of explanations of the Willis Process, and on the job training with his consultants until evaluators become comfortable using the Willis Plan. Willis' approach is mostly "learned by doing". Training usually spans a two week period, and

towards the end of the first day or maybe into the second day, Willis distributes a questionnaire and has the group go through an evaluation exercise. Willis instructs evaluators not to make assumptions about the work, and to look for facts when completing the questionnaire.

316. Willis trains his own consultants in the Willis Process. Part of their training is directed at attitudinal problems relating to stereotypical work. This part of the training, with both consultants and committees, is informal. The perspective he conveys is to ignore whether a job is male-dominated or female-dominated. He discusses attitudes with his consultants and trains them to deal with attitudes in terms of examining pieces and components of a job, breaking a job down into a number of parts and examining the pieces without regard to the sex of the incumbent. The same method is then imparted by his consultants in training evaluation committees.

317. Willis was asked to comment on a publication of the Ontario Pay Equity Commission (Exhibit PSAC-71), a commission established to assist in the implementation and administration of the Pay Equity Act (Ontario). The publication contained information on training job evaluation committees. Willis agreed in principle with the Ontario Pay Equity Commission's list of elements of appropriate training for an evaluation committee which include: information on the history of job evaluation; how salaries and wages were set in the past; pay equity and wage determination processes; how gender bias may enter into evaluation systems; trends in women's participation in the labour force; the rationale for pay equity; and specific mechanics of the system used by the organization in question.

318. However, Willis prefers his approach which, over the past 20 years, has been more pragmatic than the detailed criteria listed by the Ontario Pay Equity Commission. He says the following in Volume 209, at p. 27088, line 23 to p. 27089, line 16:

> A. We have found through experience that the best way of dealing with differences in different kinds of jobs -- and, incidentally, there is no such thing as an all women's job or an all men's job any more. They are all some mix of men and women and there are all kinds of jobs.
>
> There are some features in men's jobs and women's jobs that are somewhat hidden and there is such a variety of kinds of jobs, particularly in the public sector, that our experience has been that we can best deal with it if we don't try to focus on men's work versus women's work at all, but, rather, focus on breaking the job down into factors and examining those factors without regard to whether it is a woman's job or a man's job, making sure

that all of the hidden elements, whatever they are, are brought out.

319. Willis' own training in gender sensitization was not achieved through any formal program but rather "came from the school of hard knocks." (Volume 209, p. 27168).

320. On further cross-examination by Complainant Counsel, Willis agreed consciousness raising or gender sensitization "is not a bad thing". In response to expert evidence from Weiner and Armstrong on this topic, advocating the kind of training recommended by the Ontario Pay Equity Commission, Willis, relying on his experience, said it may be helpful to have sensitivity training of this type, but that it is not absolutely necessary. More particularly, he confirms this in his response in Volume 209, at p. 27096, lines 9 - 17:

> Q. So you are satisfied, then, that you can do a pay equity study and do fair evaluations of jobs without the kind of training that is suggested by Dr. Armstrong and Dr. Weiner?
>
> A. I would say that we have ample experience in evaluating male and female jobs in cases where there has been no sensitivity training per se, but that the consultant's guidance is sufficient.

321. The mechanism or safeguard Willis uses to ensure sound, reliable results, in the absence of more formal gender sensitivity training, is one of consultant participation. With the exception of three weeks, Willis personally observed the work of the MEC. During his three week absence, he was replaced by his consultant, Drury.

322. There were five consultants working on this study, including Willis. When the first five evaluation committees began their evaluations, each committee had a designated consultant for training and consultation. (Volume 60, p. 7433).

323. Willis testified the role of the consultant is to evaluate privately while the evaluation committee is doing its evaluations. Willis said initially the MEC would have a short period of time to discuss the particular job selected for evaluation. As was his usual practice, Willis would have his own list of questions which needed to be answered with regard to a particular questionnaire. If this information was not brought out by the MEC members then Willis would raise these questions himself.

This function was performed by his consultants with the later multiple evaluation committees.

324. Willis summarizes the role of the consultant as generally to serve as a group facilitator, to be a trainer, to answer the committee's questions about evaluation techniques, and at the same time, to observe the functioning of the committee and to maintain a finger on the pulse. When the consultants do their own evaluation of the job, they do not communicate to the committee the result of their evaluation. The purpose of these evaluations is to enable the consultant to track the committee evaluations. In Willis' opinion, the consultants have two advantages over the committees: (i) they are professional evaluators; and (ii) they do not carry any baggage.

325. In Willis' opinion, a disadvantage of the consultant's role is that as "outsiders" they do not know the environment of the organization as well as the evaluation committee members and thus, do not know how differences are perceived within an organization. Willis also points out there is always the danger a consultant may be influenced by their knowledge of a job in another organization which may be similar but not exactly the same as a job within the study.

326. The consultant is not only examining the factual basis each evaluator is using to justify their own evaluation, but is also examining what the committee is doing and more importantly ascertaining the committee's rationale for what they are doing for each of the subfactors in the Willis Plan. The consultants exercise what Willis describes as an "empirical judgment" during this process.

327. Willis testified the MEC was a good and effective committee. Based on his own observation and the information received from Drury, he was satisfied with the degree of consistency the MEC had in terms of its own discipline. His overall assessment of the quality of their efforts was they were evaluating based on facts.

328. From his personal observation, he identified two individuals who seemed to be outliers. The MEC did not tend to be influenced by these individuals. Willis' conclusions on how the other members of the MEC received and reacted to the outliers' comments was based on the remaining evaluators' reasons for evaluations and the overall consensus of the group, which were not affected by the outliers.

329. With respect to the training of the multiple evaluation committees, Willis found at the conclusion of the first five days training, the majority of the members were barely comfortable with the system, but became more comfortable with the plan after two weeks of training. Individual evaluators who testified at this hearing also experienced an increase in comfort with the Plan as their work progressed.

330. Willis recognized the need for constant vigilance in maintaining and understanding the plan so that evaluators would not revert to previous evaluation judgments. As a result, Willis met regularly with the initial five evaluation committees to review problems and suggest solutions. In addition, the Willis firm prepared technical advisories, written explanations by the consultants, which answered questions posed by the committees concerning the interpretation on the technical aspects of the Willis Plan.

(v). Master Evaluation Committee's Evaluations

331. Willis provided the Tribunal with his conclusions regarding the work of the MEC during the JUMI Study. He repeatedly maintained the MEC's evaluations were unbiased, he was comfortable with the MEC's work, the information the MEC was using was based on facts, and, in his opinion the MEC had done an excellent job.

332. On several occasions, throughout the JUMI Study, Willis was asked to assess the quality of the MEC's ratings. In response to Respondent Counsel, Willis commented in Volume 75, at p. 9202, lines 14 - 23, as follows:

> I probably examined those 503 evaluations and the differences to death. It was over a six month period that I was continually challenged about them. Every time I reviewed them by myself and with other consultants, we came up to the consistent opinion that, while there was some differences, the Master Evaluation Committee's benchmarks were satisfactory, recognizing that this is not an exact science; it is an art.

333. The approach adopted by Willis to validate evaluation results was to personally, or have one of his consultants, re-evaluate selected questionnaires. The first testing of the MEC evaluations was conducted by Willis consultant, Drury, in the spring of 1988. Willis testified the purpose of Drury's review was not to validate the results of the work of

the MEC, but to ensure the MEC evaluators knew how to use the Willis Plan and that they understood and were interpreting it properly.

334. Willis was interested in whether the MEC evaluators were consistent among themselves. The MEC evaluators themselves wished to have a review as a double check while they were still learning the evaluation process. It was not made known to the Tribunal exactly how many MEC questionnaires Drury reviewed. Her review was done in the spring of 1988, and the MEC had been evaluating since the fall of 1987.

335. For purposes of her review, Drury used only those questionnaires done when she was absent from the MEC discussions. In the end, she identified a total of 12 positions which she evaluated slightly different from the MEC's. Drury's differences arose only with female-dominated positions and the female evaluators on the MEC took exception to this fact and wrote a letter in protest. This letter was addressed to Willis and the Commission. Willis believes this controversy arose not so much because Drury was critical of the MEC's evaluations but rather because there was an appearance of singling out female jobs.

336. Willis reviewed the 12 evaluations Drury identified. He considered Drury's assessments "sound" and communicated his findings to the Commission. Of the total, in three of the twelve questionnaires there was less than a 2.5 per cent difference in points between Drury and the MEC. Drury then met with the committee once more. The MEC made some changes in view of Drury's re-evaluations, but seven out of the twelve were left unchanged. Of these seven, Drury deemed the MEC had undervalued two and overvalued five. Willis was not concerned with the small number of differences per se, he was more concerned with the fact that of the seven, five were in one direction and two in another. This fact lead him to consider the possibility of a pattern of bias.

337. Willis again met with Drury and with the MEC. Of the five jobs Drury deemed over-evaluated, four were nursing positions. After discussing the content of these jobs with the MEC, Willis concluded the MEC's

evaluations were satisfactory and he supported their original evaluations. In discussing this situation later with Drury, she admitted to Willis her past experience with nursing positions had been in the State of Connecticut and this had coloured her evaluations. Willis reasoned that nurses in Connecticut do not possess the same breadth of knowledge required from nurses in the Canadian system. Thus, in Willis' opinion, Drury was probably "a little off".

338. In his final analysis of the re-evaluations, Willis did not believe the MEC had over-reacted to the Drury re-evaluations in any systematic way. He expressed this opinion in a letter to the management side co-chair of the JUMI Committee, Lise Ouimet. The letter was written on December 5, 1988 and reads in part:

> In the spring of 1988, we responded to a request from MEC to review and comment on the evaluations they had completed as of that time. Jan Drury reviewed the Committee's efforts and made recommendations regarding twelve evaluations. Of these, four were for total point adjustments of between 10.0 percent and 10.9 percent, four were between 11.0 percent and 15 percent and one was slightly greater than 15 percent.
>
> The group reviewed her evaluations and explanations, both written and verbal, and changed their evaluations of two of the nine positions (including the one that showed a difference of slightly of 15 percent), leaving seven that are different from Jan Drury's evaluations by between 10 percent and 15 percent. Of these seven, Ms. Drury's evaluations were higher on two positions and lower on five positions. Five of these seven are nursing related positions.
>
> Comments by MEC members indicated that they believed there is a slight difference in the roles of Government of Canada nursing positions having specialty assignments than Ms. Drury's experience with nurses in the U.S. would suggest. For example, MEC gave more weight to #83 Staff Nurse-Sexual Offenders Unit's role in counselling of offenders than Ms. Drury did.
>
> I am not inclined to totally discount the MEC's judgment on this issue without more information and do not feel that these slight differences warrant concern. On the other hand, if you disagree, I suggest that these nursing positions be submitted to MEC for review including obtaining additional information regarding the significance of specialty assignments, and re-evaluated. (Exhibit R-35)

339. Another procedural safeguard designed by Willis to address the issue of disagreements arising between the multiple evaluation committees' and the MEC was to permit and even encourage the multiple evaluation committees to submit their differences with explanations to the JUMI Committee. Willis had proposed in his plan that the MEC be reconvened to address these differences, and to either explain their evaluations in a more comprehensive manner or to adjust their evaluations to conform with

the results of the multiple evaluation committees. He believes this is a vitally important exercise.

340. Willis explained once a committee has evaluated a number of jobs, they develop a sense of confidence in their own ability to evaluate and inevitably there will be minor differences in how a job is perceived. Willis said one approach would be to tell evaluation committees they would have to adopt the discipline of the MEC regardless of any disagreement. But Willis desired more open communication. If the multiple evaluation committees were not comfortable with an evaluation by the MEC, Willis felt they had an obligation and a right to note these differences, and that the MEC should review any challenges brought forward by the evaluation committees.

341. Pursuant to the above, a total of 48 challenges to the MEC evaluations were brought forward by the multiple evaluation committees. There was disagreement within the JUMI Committee as to whether or not the MEC should be reconvened to review these evaluations. At one stage, the consultants were asked to independently review approximately 33 adjustments suggested by the evaluation committees. Willis testified in two-thirds of the cases, the differences were so nominal, that they were hardly worth considering, and those cases were discussed individually with the evaluation committees. Willis believed there were about 14 remaining questionnaire evaluations requiring review by the MEC.

342. Willis did not wish to say what the change ought to be because he did not want to "second guess" the MEC. There were 14 he identified as "problematic questionnaires", suggesting a possible problem or at least enough doubt that they ought to be revisited.

343. Willis felt strongly the MEC should be reconvened to put to rest the differences in interpretation between the MEC and the multiple evaluation committees. It was ultimately decided by the JUMI Committee that the MEC would not be reconvened, instead a smaller version of the MEC (the "Mini-MEC") was created. The Mini-MEC was composed of a small number of former MEC members. They were three in total, Willis, Joanne Labine, a union representative, and Michel Cloutier, a management representative. These latter members were previously identified as "outliers" on the MEC. Both outliers, Labine and Cloutier, had been identified by Willis in his direct observation of the MEC and that conclusion had been confirmed by statistical analysis conducted by an independent statistician. It is one of the incomprehensible decisions of the JUMI Committee.

344. Not surprisingly, Willis questioned the JUMI Committee's decision to select those two outliers and suggested choosing two other individuals. According to Willis, "they stone-walled me" and he lacked the authority to overrule the JUMI Committee's decision. He felt the two outliers were ill prepared to represent the JUMI Committee because of gender bias in their original evaluations.

345. The Mini-MEC considered the consultant's recommended changes to the challenged benchmarks. The union representative agreed with the

consultant and the management representative rejected all of Willis' recommendations.

346. The Mini-MEC then suggested three options to the JUMI Committee which were made without Willis' consultation. These options included:

OPTION 1

> It is proposed that MINI-MEC review the consultants' recommended changes to MEC bench marks (33).
>
> -Should the two MINI-MEC members agree, the challenged bench marks and rationales will be amended.
>
> -Should MINI-MEC after consultation with N.D. Willis or Jane [sic] Drury can not arrive at a decision, than MINI-MEC and the consultant will determine whether it is in the best interest of the study to remove the bench mark(s) in question.

OPTION 2

> It is proposed that challenged MEC bench marks not be amended. In cases where MINI-MEC agrees that the rating is an inconsistent one, then the bench mark(s) in question(s) [sic] would be removed.
>
> This proposal is based in part on our opinion that it is to[o] late to attempt to change bench marks and have the committees adjust their rating patterns.

It is to be noted that the two above options would necessitate re-sore thumbing in situations where a change to or the removal of a MEC bench mark is effected.

OPTION 3

It is proposed that should a situation arise where a committee is unable to reach a consensus on a rating, that the questionnaire be referred to MINI-MEC for resolution.

It is also recommended that further challenges to MEC bench marks not be accepted.

347. The JUMI Committee selected Option 2. As might be expected the Mini-MEC could not agree which benchmark rating was inconsistent and, as a result, none of the benchmarks was removed. Although disappointed, Willis did not feel the integrity of the process was invalidated. At this stage he believed the evaluations were "intact" and reasonable.

348. Wisner performed the re-evaluations of the challenged benchmarks and suggested an average 4.2 per cent overall increase for the 14 benchmarks under review. The purpose of the consultant's review was to

determine whether or not the differences were representative of a pattern of bias. The analysis by Willis did not demonstrate a pattern of bias, and Willis felt he could live with the JUMI Committee's decision not to recall the MEC. Willis did not believe the overall differences identified by this analysis between the consultants and the committee had a material adverse affect on the study.

349. Willis testified Wisner's analysis illustrates the percentage difference between Wisner and the MEC's approach. He stated the consultants tried throughout the study to refrain from imposing their evaluations on the MEC. Willis was asked when he would impose the consultant's evaluations on the committees. He responded in Volume 57, at p. 7053, line 14 to p. 7055, line 10:

THE CHAIRPERSON: Maybe you can tell us: Where do you draw the line between when you strongly make a recommendation or you strongly suggest or you advise? Where do you draw the line in terms of saying to the Committee, "This should be done", or do you ever do that?

THE WITNESS: Yes. First of all, the consultant who is meeting or sitting with the committee will be privy to the questionnaire information and the discussion about the job. If we find at that time that people are not talking about the facts and are apparently not using the facts fairly and equitably, we would raise the question with the Evaluation Committee itself as it is proceeding.

On the other hand, after the fact, looking at a series of evaluations, we might disagree slightly with the committee, but our concern would be whether or not that difference might be a difference of honest interpretation by the committee, it might be a difference between what the consultant knows about that kind of a job -- and we can be a little bit misled ourselves as consultants -- as opposed to the committee, which may have a better handle or feel for the content of jobs in their organization.

If we identified a pattern that seemed to be resulting, then we would take very strong steps. Recognizing that these are value judgments, we have to have some tolerance. Just because we come out with an average of five or six per cent more overall than the committee, that doesn't necessarily mean that they are five or six per cent wrong. But our concern would be more: Is there a pattern here or is there a random difference? If it is a random difference, then we are not at all concerned, unless there is a possibility that they are misunderstanding how to use the instrument itself.

THE CHAIRPERSON: Why were you strongly advising that MEC reconvene?

THE WITNESS: Just because of the psychology of the committees, they felt very strongly about this. Even though there may be a very slight difference in a job, they feel uncomfortable if they haven't at least had a hearing.

350. According to Willis, stemming from the JUMI Committee's decision not to reconvene the MEC, a considerable amount of frustration was experienced by the multiple evaluation committee evaluators. The consultants were obliged to tell the evaluators the JUMI Committee had made a policy decision and that there would be no changes in the benchmarks resulting from the committee challenges. Willis suggested that the

committees try to work around them. He believes many evaluation committees tended to take alternate MEC evaluations for comparison with their own evaluations and tended to ignore the MEC evaluations which had been challenged.

351. Willis indicated the level of frustration was highest when the evaluation committees expected and were waiting for the MEC to reconvene. When they were informed this was not going to happen they became more resigned.

(vi).Multiple Evaluation Committees' Evaluations

352. Some of the first five evaluation committees tended to negotiate rather than cooperate in trying to achieve a consensus. Willis found the evaluation committees tended to balance each other fairly well, but the obvious result was lower productivity. In the initial formation of the five evaluation committees, Committee #3 was more contentious and less productive than the others. Willis trained Committee #3 and led them for the first three weeks of their evaluations. He met with the chair weekly to try to work with her, as he considered her part of the problem. He sat with this committee a great deal of the time, working with them and monitoring their evaluations. He was better acquainted with this committee and their problems than with any other committee. Willis observed Committee #3 had individuals on the staff side who seemed to be committed to rating jobs from female-dominated occupational groups as high as they could and two or three on the management side, some of whom had former classification backgrounds, who were devoted to keeping the jobs as much in line as they could. He described this as a "stand-off".

353. Willis testified the chair of Committee #3, a union representative, sometimes assumed the role of evaluator and entered into the discussions in a way he believed was inappropriate for the chair. The proper role of a chair is to assume a neutral posture and to facilitate committee discussions. Willis subsequently recommended to the JUMI Committee that the chair of Committee #3 be removed and he eventually recommended the entire committee be disbanded. However, the JUMI Committee rejected his recommendations and nothing happened to improve Committee #3's situation until the reformation and expansion into the nine committees.

354. Willis described a good functioning evaluation committee as a team working together with each member trying to evaluate fairly and equitably. His discomfort with Committee #3 was not the fact of the actual

evaluation ratings but rather the manner in which they evaluated. This committee would debate until finally they would agree due to exhaustion.

355. Willis did not believe the "stand-off" he described between management and union evaluators on Committee #3 negatively affected the evaluation process in that committee. Neither he nor his consultants could detect any pattern of bias in Committee #3 evaluations. However, as a consequence of the "standoff", some committee members experienced health problems and the productivity rate suffered noticeably.

356. Willis also had a problem with the initial formation of Committee #4. He testified Committee #4 was an excellent committee from its inception to about March of 1989. However, in the latter stages, due to substitutions and the reforming of this committee, problems developed. In April of 1989, Willis requested Committee #4 undergo a final sore-thumbing exercise. During this exercise the chair of this committee came to him, almost in tears. Willis testified she said, "I can't handle this any more. It has all broken down, they are all getting emotional, they are yelling at each other. We have a job to do and I quit." In the JUMI Committee minutes of October 31, 1989 (Exhibit R-44), Willis remarked, with regard to the consultant report on Committee #4, the "major problem with Committee #4 was its lack of objectivity, creating the disastrous consequence of two camps, separate agendas, and arbitrary and opposing viewpoints."

357. At this point the committee had evaluated 52 jobs. Willis then requested that the remaining committee members state in writing their individual concerns about the evaluations and suggest any changes which they thought were necessary. He then disbanded the committee. Subsequently, a Willis consultant, Robert Barbeau, reviewed the specified concerns, made recommendations and was asked to take appropriate action. The committee members made suggestions on a total of 25 jobs and there was only one in which the consultant differed significantly from the committee members. Willis described this as one instance where there was consultant influence on the evaluations, albeit a small amount.

358. Willis did not observe any problems occurring with Committee #1 or #2 during the initial formation of the five evaluation committees.

359. Willis' observations of Committee #5 were that the evaluators tended to be extreme, on one side or the other, but not as extreme as Committee #3 and their productivity "tended to move along". Willis identified one female union representative demonstrating a female preference and two male management representatives demonstrating a male preference. Willis further testified a female union representative also demonstrated a male preference. Willis found two of these evaluators, a female union representative and one of the male management representatives,

tended to cancel each other out. Willis observed that the other members of the committee were not influenced by these two individuals and tended to discount their positions.

360. Willis further found the evaluations generally produced by Committee #5 to be "pretty good". He identified two members of the committee as outliers but later recommended they become chairs of the

expanded nine committees notwithstanding, because he considered them to be good evaluators.

361. The Tribunal heard evidence from three evaluators who were members of Committee #5. Each confirmed this committee's thoroughness in discussing jobs and diligence in completing their task. Their evidence further corroborates Willis' view that the outliers did not influence the consensus of the committee.

362. There was evidence provided by two evaluators who were members of the first version of Committee #5, to the effect that the questionnaires discussed in this committee were difficult. One of these evaluators, Mary Crich explained the committee's long discussions resulted from very difficult male jobs.

363. Pauline Latour, another evaluator on Committee #5, states in Volume 171, at p. 21604, lines 20 - 25:

> A. We had a difficult -- the questionnaires in Committee 5, I have a sense that they were more difficult to evaluate. There were many that we seemed to have unanswered questions. So, we definitely returned more questionnaires in Committee 5.

364. Latour further elaborates on this point in Volume 171, p. 21605, line 12 to p. 21606, line 9:

> Q. You mentioned just a short time ago that there were some jobs that you recall as being more difficult to evaluate than others. Could you describe which of these -- give us perhaps some examples of the types of jobs that as a committee you found more challenging than others.
>
> A. This perhaps is going to be a bit of a convoluted answer, but, for example, the jobs that we were comfortable with were jobs that we had rated many similar positions. For example, we

evaluated many secretarial jobs which were evaluated at quite a range, from typists to senior executives. We had a good understanding of the nature of the work.

There were some positions where we evaluated basically one or two jobs that were related and we never had a sense of how that job actually fit in the section that that person worked in. So, because they were so unrelated, there were quite a few positions that were unrelated, we really had a difficult time just grasping the level of complexity of that position.

365. The Tribunal heard direct evidence from 15 witnesses, who were evaluators on Committees #1, #2, #3, #4, #5, #8 and #9, about their experiences and perceptions while serving on their respective committees. Evidence about Committees #6 and #7 was provided by Willis and another Willis consultant, Owen. Neither one expressed any serious concern about what they observed on either of these committees.

366. In terms of the direct evidence from these 15 witnesses, the Tribunal was impressed with their individual level of commitment to the study. Although job evaluation is a systematic process that is mentally challenging, the fact remains these individuals endeavoured to achieve a consensus evaluation for each position, eight hours a day, five days a week, over long periods of time. Willis observed variations in the productivity of the committees. The productivity record based on a total of 3,185 questionnaires is as follows: (i) Committee #1 - 466; (ii) Committee #2 - 431; (iii) Committee #3 before expansion to 9 committees - 165; Committee #4 - first version - 200 evaluations; second version - 52 evaluations and after expansion of 9 committees - 160 evaluations; Committee #5 - 430 evaluations. After the expansion to 9 committees, Committee #6 - 197 evaluations, Committee #7 - 149 evaluations, this was francophone committee; Committee #8 - 150 evaluations, this also was a francophone committee and Committee #9 - 145 evaluations.

367. Given his experience in previous studies, Willis expects a certain amount of conflict within an evaluation committee because of the different backgrounds and perspectives of the various evaluators. However, Willis testified the degree and nature of the conflict he observed in this study within the evaluation committees made him feel uncomfortable.

368. Some of the problems which arose in the multiple evaluation committees had been anticipated by the JUMI Committee. The Testing Sub-Committee of the Willis Evaluation Plan, in its report of July 20, 1987

(Exhibit HR-11A, Tab 19), made recommendations in response to problems experienced by this sub-committee during a two week trial period. Some of the problems included personality conflicts, weariness owing to constant concentration and stress in being seconded from their regular jobs for long periods of time. As a consequence of this experience, the sub-committee recommended the rotation of evaluation committee members between evaluation committees, working a shorter day or week and the utilization of alternate members to replace designated committee members for periods of time. These recommendations were never acted upon by the JUMI Committee and the Tribunal was not provided with reasons for the rejection of these recommendations.

369. The evaluators testified they experienced tension as committee members, stress in reaching a consensus, personality conflicts, inflexibility on the part of some individual evaluators, difficulties with some chairpersons and screaming by some evaluators. In some instances evaluators walked out of evaluation meetings because of frustration. Compounding these problems was the frequent rate of substitutions of members for some committees. This resulted in a change of dynamics requiring adjustments by both new and older members.

370. Coupled with these problems, was a rigid working environment orchestrated and controlled by the Chief of the EPSS, who was apparently more flexible with management evaluators than union evaluators. The Chief, Pierre Collard, closely monitored the arrival and departure time of the evaluators, the lunch breaks and the coffee breaks. He insisted doors remain closed at all times during deliberations (causing ventilation problems), limited access to telephones, and kept all supplies in locked

compartments (thus creating time delays for obtaining supplies). These very stringent constraints intensified the frustrations already experienced by committee members. Moreover, some evaluators were "from out of province" and found it difficult to wait for long periods to be reimbursed for their travel expenses. This issue, in particular, was not resolved in a timely fashion.

371. Many evaluators who testified at this hearing, expressed a willingness for and the necessity of adhering to the MEC benchmarks as well as to the requirement that evaluations were to be based upon facts contained in the questionnaires, and not on any other extraneous considerations.

372. The only criticism the Tribunal heard concerning the committees' willingness to follow the MEC discipline was that for a short period of time, early in the evaluation process, Committee #1 tended to follow its own discipline rather than that of the MEC. This problem was corrected as soon as it had been identified by the Willis consultants.

373. In the early part of the year, 1989, Willis began to express to the JUMI Committee, his concerns, not directly related to the actual evaluations themselves, but concerns regarding "circumstantial things" which had transpired. He referred to these incidents as "smoke" because they were largely rumours and included incidents which occurred both inside and outside the committee rooms. He became increasingly uncomfortable with how the evaluation committees were working and with what he described as confrontations between union and management sides. Although he could not identify anything specific which would suggest gender bias was developing, based on his own observations and those of his consultants, he knew "some things were happening" and some improper attitudes were developing causing him a great deal of concern.

374. On several occasions, while Willis sat with a committee, it became clear to him a position taken by a particular evaluator was very biased. Usually, the individual evaluator refused to change the score, even though lacking the facts upon which to base a rating. The frequency of these occasions began to disturb Willis. It was occurring, he observed, on both union and management sides, and arose more frequently from the earlier formation of the five evaluation committees and their members.

375. Willis said that he was made aware of union members attempting to recruit other evaluators to their "bloc". He had not seen this phenomenon in any other evaluation study in which he has been involved. Willis did not observe directly any incidents regarding this recruitment. He was informed, however, by Owen, of an incident in which an evaluator approached another evaluator about the evaluations. Owen testified about the circumstances surrounding this incident which occurred in February, 1989. Owen testified he overheard a conversation between two female evaluators who entered a room in which he was working. He overheard one female evaluator say to the other "we don't think you're doing enough for women's jobs." According to Owen, the other evaluator became agitated, her voice increased in loudness and he heard her reply "I didn't come here to build

up some kinds of jobs. I came here to do an honest job of evaluating the work."

376. Owen further testified he observed a sort of "faction-based" behaviour in the committees. There were some union evaluators who seemed to be treating certain jobs in a similar way as union evaluators in other committees. He identified them as Alliance members. What troubled Owen was in his prior experiences, which involved training and facilitating more than 50 evaluation committees, he had not observed any kind of similar behaviour. He also noticed unusual scoring, long discussions advocating a particular choice, and the selection of benchmarks inappropriate to the particular evaluation at hand. In another incident, during the initial formation of the five evaluation committees, Owen was asked to chair Committee #3, because the regular chair was participating as an evaluator elsewhere. When the chair returned to the room, a very contentious argument concerning an evaluation was taking place. The chair asked Owen to rule on how to proceed and asked for points of order similar to Roberts Rules of Order. Owen was completely unfamiliar with Roberts Rules of Order and was thus unable to give an appropriate response. The chair's reaction was to order and instruct the Alliance evaluators on this committee to walk out, which they did, slamming the door as they left. Owen viewed this unhappy incident as an attempt by one side to control that particular committee.

377. Like Willis, Owen felt frustration at not having any level of opportunity to intervene or take action.

378. Another incident noted by Owen occurred during the fall of 1988. Most of the Alliance members did not attend their committees on a particular day as they had designated it the day for a "sick out" to demonstrate their support for pay equity issues at the collective bargaining table. Apparently, collective bargaining was under way and there was some discussion among union members as to whether the proposals on pay equity would be withdrawn from the bargaining process. Two Alliance members who did not attend the sick out, told Owen that they were concerned about reprisals from their union for not having participated in the sick out.

379. Among the committees, Willis felt the conflict was too much "us versus them". Willis confirmed he had never seen so many participants with a classification background in a pay equity study and this was "an important aspect in the conflict in this case."

380. Willis testified if the Federal Public Service was his organization and he had control over the evaluation process and decision making authority, he would have made some changes and continued with the study. His preference would have been to remove the personnel creating problems and engage more consultants to work closely with each committee.
381. Willis' expert opinion is that gender bias can operate very

subtly in a pay equity study, and he felt in order to defend the results, he had to reassure himself there were no problems with the evaluations. Willis was not sure the actual problems which existed resulted in biased results. He stated in Volume 69, at p. 8654, lines 8 - 14:

> I have mentioned that there was an interesting contradiction. I had some very strong concerns about attitudes, things we observed. However, when we attempted to look at what committees' results were and when we tried to look at comparison of similar jobs, we were not able to detect a clear pattern of a problem.

382. During the evaluation of the first five evaluation committees, Willis testified that based on his observations, he identified ten evaluators who he believed were exhibiting gender preferences. According to Willis, the majority were exhibiting a female preference. His approach in dealing with this problem was to counsel these individuals. At this stage, Willis could not determine whether the identified evaluators were influencing the group evaluations. He was concerned he had no evidence other than his personal observations for support. He alerted his consultants who were already aware of the particular individuals. He and his consultants continued to track these individuals and to look at the results of the evaluations overall.

383. Another approach of the consultants was to break the evaluations down by occupational groups and determine if these individuals were influencing the group and to then attempt to identify on an overall basis if there appeared to be any problems with bias. This tracking did not seem to indicate any significant bias.

384. When counselling individual evaluators, Willis would sit with them in a private room and discuss their evaluations and what changes he expected from them. Willis testified he did not see any difference or change in the evaluations of the individuals after they received counselling. Willis was informed during counselling of some management evaluators that they were evaluating to offset evaluations on the part of the union evaluators. For the most part, Willis did not receive denials from any of the evaluators whom he counselled as to their behaviour.

385. Throughout the study, Willis also conducted committee counselling. He observed an evaluation committee as they were evaluating. In his interventions, he attempted to direct the evaluators to the facts, to look at the questionnaire and discuss the actual position rather than to make assumptions or stereotype. As to the effect of committee counselling

Willis said the following in Volume 57 at p. 7087, line 9 to p. 7088, line 5:

> Q. With regard to the last type of counselling you just gave for the evaluation committees as a group. I had already asked you as to your opinion of the efficacy for the individuals. Now I want to know what your opinion is with regard to how well the counselling of the evaluation committee groups worked?
>
> A. That is a little hard to say. These committees were somewhat unusual compared to most committees I work with, in that I was not observing actual evaluation bias or any pattern that I could identify. On the other hand, I did not have committees that were all working together to accomplish a fair, equitable, conscientious result.

> What I had in many committees were the staff on one side and the management on the other side and they were at loggerheads. This was a pattern that was not universal, but we found it on several committees. The extent to which our counselling affected them, in some cases, was negligible. [emphasis added]

386. During later testimony, Willis was asked to explain what he meant in the above excerpt by the words "I did not have committees that were all working together to accomplish a fair, equitable, conscientious result." Willis explained his reference is primarily to the word "conscientious". To him this word suggests an employee is working hard and meeting their own personal standards. In this context, Willis testified every individual he observed on every committee was evaluating conscientiously. On the other hand, the consultants attempted to instill a standard by which every job would be treated fairly, objectively and impartially. In that context, Willis said he observed evidence, which was not pervasive among all evaluators and committees, that this standard was not being consistently applied.

387. The testimony from the participating evaluators who were asked about how they personally approached evaluations was that they were honest, dedicated and conscientious. They observed the same commitment from most of their committee members.

388. Specific questions were posed to the evaluators who testified about Willis' concerns, referred to as "smoke". The questions posed concerned rumours some committees were "block voting", meaning union

evaluators would vote together to obtain the same score for subfactors and all of the management evaluators would vote together to obtain their same score and about other methods of communication including the use of sign language and hand signals to indicate how specific evaluators were scoring so as to influence decisions.

389. None of the evaluators who testified observed this kind of behaviour or any other kind of organized communication designed to over-evaluate female jobs and under-evaluate male jobs. Apparently, hand signals had been discussed in a social setting, which one witness believed resulted from frustrations expressed about the difficult process of job evaluation. This action was given and received in a joking manner.

390. The Tribunal heard direct evidence of three separate incidents of inappropriate behaviour. In the first incident, both evaluators were female representatives of the Alliance who were involved in the conversation overheard by Owen, referred to earlier. An evaluator on Committee #4 testified she was approached by another evaluator on her committee concerning the subject of whether or not she was evaluating female-dominated jobs fairly. This witness had the impression that this individual wanted her to increase her ratings. The witness testified she responded by saying she was there to evaluate fairly and to the best of her ability in comparison to all of the jobs. As far as the witness was concerned, this was the end of the incident.

391. The second incident also occurred between two female Alliance evaluators. The witness testified she was approached by another evaluator who wanted to meet her outside of the committee room to discuss how to evaluate jobs. The gist of the meeting was the second evaluator wanted the first evaluator to favour female-dominated jobs in a higher bracket in the same way as she did. The first evaluator felt this was not an objective approach and told the second evaluator that her ratings would continue to be objective.

392. With regard to the first and second incidents respectively, both witnesses testified the incident did not have any impact on their manner of evaluating. The evidence is clear the individual connected to the first incident and who made the request was noted by her committee for her biased ratings which the committee had endeavoured unsuccessfully to change. Since she refused to change, she was basically ignored by the rest of her committee.

393. The third incident involved a female, Institute evaluator. This evaluator testified there was a social gathering in her hotel room involving about 10 or 15 evaluators. A conversation occurred later in the evening between this evaluator and four other evaluators from the Alliance. The Institute evaluator testified she had been advocating an objective point of view for doing evaluations and two of the Alliance members became very aggressive toward her. Their response was the study was an opportunity for women to have something done for them, and nothing was going to get done unless women's jobs were evaluated higher and the study was their last chance. The Institute evaluator testified "things then got a little too personal." Another Alliance witness who testified described this incident as a verbal attack on the Institute evaluator.

394. With regard to this third incident, the Institute evaluator assumed the individuals who confronted her in her hotel room were in a position of authority vis a vis the Alliance and could call meetings and influence other Alliance evaluators. At the time of giving her testimony, she admitted she no longer had any basis for this belief and no longer felt there existed a common understanding among Alliance evaluators to act dishonestly.

395. Willis recalls he had discussions about problems in the evaluation committees with the Mini-JUMI, a sub-committee of the JUMI Committee. This sub-committee was formed to handle procedural problems of the evaluation committees. Willis testified he discussed with two of its members, Gaston Poiré and Elizabeth Millar some of the evaluators he felt were creating problems. Willis suggested certain individuals be eliminated from the evaluation committees. He testified he did not get the active support he expected. As a result, the JUMI Committee reassigned problem individuals when the committees expanded from five to nine. According to Willis, after the committees expanded, some committees worked well and some still had problems but not to the same extent as the initial five evaluation committees. He stated "Nothing was worse than the original Committee #3." In his estimation, it was at the bottom of the barrel and after that it "got better." (Volume 69, p. 8653).

396. Willis regarded what was happening in the evaluation committees as unacceptable. He concluded he needed to conduct further analysis, a more in depth analysis of the results, if he was going to be able to support the outcome of the study. Although he had not identified gender bias in the evaluations by January and February of 1989, he said in Volume 58, at p. 7229, line 13 to p. 7230, line 3:

A. I think the one thing that characterized the whole study, the equal pay for work of equal value charge, was to evaluate a broad range of positions on a gender neutral basis. I think everything we did in terms of the process that we set up and the evaluation system that was used, the way we tried to work with the groups, was all aimed primarily at avoiding any evaluations that would suggest traditional relationships, or in any way any bias that could be identified as gender bias.

I feel that at all stages in this study it was paramount that we continue vigilance and continually reinforce the need for objective, fair, equitable evaluations of any and all kinds of positions.

397. A letter dated May 4, 1989 to the JUMI Committee co-chairs from Scott Gruber, a Willis consultant, contained a recommendation that a special analysis of evaluation committee results be undertaken. The letter reads in part:

This letter describes our proposal for a special analysis of evaluation committee results, which we believe is timely and appropriate. The question to be addressed is:

Have the evaluations of the five evaluation committees (#1 through #5) been consistent with the evaluations generated by the MEC?

...

The methodology for this analysis will be as follows:

1. A sample would be selected randomly from the evaluation result of each committee. The sample size will be 10% of the positions evaluated, with a minimum of 25 per committee. This latter provision allows for a reasonable examination of the efforts of low productivity committees. Using these guidelines, the total sample will be approximately 140 positions.

2. A Willis consultant, familiar with the MEC evaluations, will examine each of the 140 questionnaires and make comparisons with appropriate or corresponding MEC questionnaires.

3. Based on this examination the consultant will then assess the soundness of the final, post-sorethumb consensus evaluations from the five committees together with their

selected MEC benchmark questionnaires. Problems and trends will be identified, by committee and for the entire group.

4. Gender domination information will be obtained for positions in the sample at this stage. Additional analysis will identify whether any committee, or committees, exhibited tendencies regarding male or female dominated groups in their final results. Other variables besides gender could also be included in the analysis at this stage.

5. A report will be prepared and presented to you, describing the process of the research, the analysis, and the findings.

...

We view this as a quality assurance study, to examine the evaluation results of five committees, comprised of people with a diversity of education, experience, and occupation, that could not mirror the characteristics of the composition of the MEC...A major question to be explored is whether the committees have used the MEC benchmark evaluations consistently and properly in the comparison process.

...If the results show that the five committees have performed their respective tasks consistently with the MEC, many concerns regarding the study will be resolved. On the other hand, if problems are identified corrective actions can be taken and the continuing efforts of the nine committees will benefit from the knowledge gained.

(Exhibit HR-11B, Tab 32)

398. A snapshot assessment of the validity of the evaluations was requested to be conducted on the 2,000 positions evaluated to date. Willis suggested one of his consultants examine 10 per cent of the completed evaluations and compare the committees' evaluations to the consultant's evaluations. In this way, he would at least satisfy himself there was no evidence of a problem or would expose the possibility that a problem might exist. His intention at the time was to start with a small study, which might expose evidence of discrimination. If a problem was revealed, he anticipated conducting a second study, a more in depth analysis, which would expose the extent of any problem indicated by the first study. He

did not indicate to the JUMI Committee directly that he anticipated a two-tiered approach.

399. The proposal of a small study was accepted by the JUMI Committee and this analysis commenced in the spring of 1989. The analysis is entitled the Special Analysis of Evaluation Committees' Results (the "Wisner 222") and was prepared by the Willis consultant, Jay Wisner (Exhibit PSAC-4). Wisner examined and re-evaluated 222 of the committee evaluations from both the five and nine committees. When the sample of the 222 positions was made the multiple evaluation committees were still evaluating questionnaires and the nine committees had been operating for about three months.

(vii). Re-Training of Multiple Evaluation Committees

400. This step in the Willis Process involves retraining an evaluation committee or an individual evaluator. If the consultant noticed a problem, the objective of the retraining session was to bring the committee or individual back to the MEC discipline. Retraining could be as informal as that which took place during the life of the MEC, when Willis assisted the committee in interpretation of the plan, or it could have involved more formal sessions which did occur during the work of the five and nine committees. After the initial training for the five evaluation committees during the week of September 19 - 23, 1988 (Exhibit HR-11B, Tab 27), the next formal retraining occurred in March-April, 1989, following the expansion of the multiple evaluation committees. Between these sessions, less informal training was provided by the consultants as required.

(viii). Sore-Thumbing

401. Another procedural safeguard in the Willis Process is a review process referred to as sore-thumbing which is synonymous with the term interim review. According to Willis the first interim review usually occurs after 25 to 30 jobs have been evaluated. These jobs are then listed in descending order of points and comparisons are made between the jobs, factor by factor. The idea is to look for sore-thumbs, that is to say, those evaluations which may not have the same consistency as the other evaluations. A final evaluation sore-thumbing session occurs after all the jobs have been evaluated. This technique is designed to ensure consistency within a committee and reveals whether a committee has varied from its discipline. No evaluator was permitted to be involved in a sore-thumb exercise if they had not been present during the original evaluation.

402. The MEC had five sore-thumb sessions which resulted in nominal changes. Overall, Willis was satisfied with the results of the MEC sore-thumbing. Each of the other evaluation committees also had four or five sore-thumb sessions. The evaluation committees sore-thumb exercises had a different emphasis than the MEC simply because the concern was more with whether the committees were adhering to the MEC discipline. This sore-thumbing took the form of reviewing their own evaluations and comparing them with the MEC discipline so as to ensure consistency.

403. If the evaluation committees were not consistent with the MEC discipline on a factor by factor basis, the result would be a lack of consistency in overall evaluations across the board. The degree of liberalness or conservativeness is not always the same from one factor to another. The important rule is all jobs must be treated the same way; that is, if the committees are going to be liberal in interpersonal skills, then they must be liberal with all jobs and if they are conservative in knowledge and skills, then they should be consistently conservative for this factor. Willis did not express a direct opinion on the effectiveness of the multiple evaluation committee sore-thumbing exercises.

D. RELIABILITY TESTING

404. As part of the Willis Process, Willis generally recommends reliability testing of the evaluations.

(i). Inter-Rater Reliability Testing

405. The first type of reliability testing is inter-rater reliability (IRR) testing which specifically identifies evaluators who may be developing patterns in their ratings inconsistent with the other members of their committee. Willis introduced the concept of IRR testing during the planning phase of the JUMI Study.

406. Willis explained IRR testing is advisable for two reasons. For personal reasons Willis finds, when counselling evaluators who demonstrate bias in their evaluation scores, it is helpful to have documentation of a statistical nature to support his observations and opinions. If it were otherwise, it would be the consultant's word against the evaluator's. Willis finds it helpful to use the IRR testing with the individual and to ask the evaluator to look at the pattern in their evaluations. This makes it easier to discuss the problem with the evaluator and convince the evaluator to change. He testified in certain instances, evaluators would

refuse to heed the suggestion their evaluations were biased unless confronted with statistical documentation.

407. The second reason Willis introduced the IRR testing is, in a very large and important study like the JUMI, he felt the results would be subjected to public scrutiny and in that sense, might be criticized for failing to use this procedure.

408. Willis made it clear IRR testing is not necessary in order for himself or his consultants to observe and identify outliers. Willis testified an experienced consultant will always recognize an outlier but this testing provides some written statistical evidence.

409. Willis' recommendation for IRR testing was not accepted by the JUMI Committee in the initial planning stages. He later reintroduced this concept when the MEC started its work. There was some debate within the JUMI Committee about whether or not the testing should actually be undertaken. At the January 13, 1988 meeting of the JUMI Committee (Exhibit R-9), the management side agreed in principle there was a need to conduct IRR testing in addition to inter-committee reliability testing but questioned the current Willis proposal.

410. The JUMI Committee formed a sub-committee called the Inter-Rater Reliability and Methodology Sub-Committee (the "IRR Sub-Committee") which was delegated to explore this issue. Its mandate was:

(a) to determine and make recommendations about the methodology and research necessary to test evaluation committee rater reliability

(b) to assess and make recommendations about research methodology as it applies to the JUMI Study as a whole (Exhibit HR-11A, Tab 26)

411. Willis testified he was not certain exactly why there was resistance from the JUMI Committee to IRR testing but, ultimately, it was decided by that committee to engage the consulting firm of Tristat Resources to perform the testing. For his part, Willis accepted and agreed to this arrangement. The testing was conducted by Dr. Richard Shillington, a statistician who testified as an expert at this hearing.

412. Willis was disappointed the actual IRR testing did not commence before the MEC had completed its work. As a result, he was unable to use

the results in his counselling of the MEC evaluators whom he identified as outliers. Willis stated, "but other than that, I was satisfied with the testing itself."

413. Originally, Willis had proposed to undertake the IRR testing at least three or four times during the course of the MEC's work, thus providing statistical information which he could use as a basis for discussion with evaluators who exhibited gender bias. During the MEC's work Willis identified two evaluators as outliers and the IRR testing confirmed his observation. Willis met with them but since no testing had taken place, he had no documentation to support his counselling.

414. Willis did not have authority to remove those he identified as outliers. At the time, Willis felt their biases were subtle, ineffective and not harmful to the MEC's work. In both cases, Willis' counselling had little or no effect. Willis testified the two outliers tended to cancel each other out. One was systematically favouring male jobs and the other female jobs. The IRR testing confirmed the identity of the two outliers.

415. Although, the IRR testing did not assist Willis in his effort to counsel the outliers, still, in his opinion, the testing could be used as "after the fact evidence of the consistency of the evaluation process." Shillington's report on the IRR testing of the MEC evaluations was released in July 31, 1988. The report is referred to as the "Tristat Report".

416. Shillington first became involved in the JUMI Study in the spring of 1988. He was approached by a Treasury Board member of the IRR Sub-Committee and was asked if he would be interested in the work of the sub-committee. Although Shillington was retained by the Employer, he viewed the IRR Sub-Committee as his client. In the context of the IRR testing, Shillington conducted statistical tests to analyze and interpret inter-rater reliability. Its purpose was to determine whether evaluators functioned consistently and whether evaluators treated questionnaires for male- and female-dominated occupational groups in a consistent fashion.

417. Shillington understood his role as assisting the IRR Sub-Committee to develop a methodology which could be used with the data to address their questions and to assist them in making some decisions. The IRR Sub-Committee was primarily interested in identifying evaluators who seemed to have a gender preference or a gender bias in their questionnaires but there were other aspects as well. One of these was a determination of whether these evaluators were influential evaluators within their committee. (Influential evaluators, in this context means evaluators who

seemed to be able to shift the consensus score of the committee towards their own initial rating.)

418. Shillington used a combination of statistical tests called t-tests, chi square and z-scores, (which are similar to t-tests), to make comparisons of the differences between individual evaluator scores and committee averages to determine whether there was a pattern between male and female questionnaires.

419. Shillington identified two MEC evaluators demonstrating a systematic gender preference in their ratings. One was a male management representative who allocated male-dominated positions a higher rating than the committee and the other was a female union representative who allocated female-dominated positions a higher rating than the committee. These evaluators were the same individuals identified by Willis and who ultimately became members of the Mini-MEC. The IRR test results did not indicate there was a dramatic difference between their scores and the committee scores on every single questionnaire but a rather subtle, smaller pattern appeared fairly frequently.

420. The IRR Sub-Committee had also requested Shillington identify "influential" evaluators. The sub-committee put the question: "Were there particular raters who seemed to be able to do this more often than other raters?" To answer this, Shillington looked at questionnaires where the consensus score was not near the middle of the ratings in order to determine how often particular evaluators were in the situation where they had apparently moved the consensus score towards their score. Using this methodology, some evaluators were identified as influential.

421. Shillington was then asked to identify the extent to which evaluators who had shown a gender bias were influential. These test results indicated that the evaluators who demonstrated a significant level of influence over the committee were not the same two evaluators who had been identified as having a gender bias and that the most influential evaluators displayed no gender preference.

422. The third and last aspect of the IRR Testing was the identification of questionnaires for re-review. This exercise arose from the identification of influential evaluators. The IRR Sub-Committee used the test results to identify questionnaires where the consensus score seemed to be either large or small compared to the initial ratings. Approximately 103 questionnaires were referred by the IRR Sub-Committee for re-review and characterized as "unusual". Of these questionnaires referred, one factor only, i.e., working conditions, was responsible for 43 of them being identified as "unusual".

423. Shillington testified regarding the limitations of the IRR testing methodology contained in his report of July 31, 1988. Using the methodology of comparing evaluator initial scores to committee average scores, an assumption had to be made, according to Shillington, that the committees are less biased than individual evaluators. In this context, the overall average of a committee is then considered more reliable.

424. Another limitation expressed by Shillington is found in the Tristat Report:

> Further, the fact that a rater systematically favoured occupations dominated by one gender over another does not imply a gender preference. Since the sexes were not equally distributed in the population, it may simply have been a result of a bias for or against some other factor which was common in occupations dominated by one gender. For example, a bias in favour of advanced education would have caused a rater to be identified as having a preference in favour of males having been more common in senior positions. Similarly, individual rater preferences associated with technical skills, or physical labour would have lead to the appearance of a gender bias.
> (Exhibit HR-39, p. 5)

425. As to the limitations expressed in the above excerpt found on page 5 of the Tristat Report, Shillington testified in Volume 86, at p. 10653, line 10 to p. 10656, line 11:

> THE WITNESS: The mathematical statistics are not a lot of help in that. That is basically an interpretation question.
>
> Thank you for drawing my attention to that limitation. When I was trying to summarize this report, I didn't mention it and it was an important limitation.
>
> The mathematical statistics can be helpful in identifying that an individual was treating questionnaires from male-dominated groups differently than questionnaires from female-dominated groups. But it can't do a lot to help you understand why.
>
> The limitation that is expressed in the section you pointed out, that it might be an indirect relationship to education, or blue-collar/white-collar preference, or things like that, is certainly a valid consideration. Someone who had a strong preference who

thought that advanced education was undervalued or though that work outside was undervalued or overvalued could possibly appear to have a gender preference or a gender bias -- I will use those words interchangeably for a moment -- and you would have no way of knowing whether or not it was directly related to gender or an indirect relationship to something that is correlated with gender.

If that idea that I discussed of having hypothetical questionnaires inserted into the process, questionnaires that were basically rigged to appear to have a gender difference even though they were identical in all other respects, if you had done that, then you could have actually addressed some of this concern.

As part of that you could have said: If we had someone who had a high education preference and that high education preference might get reflected in gender bias, can we design three or four questionnaires which are all similar in terms of requiring advance

qualifications, but are different in terms of a male/female composition? Then you could look at those questionnaires for these individuals and try to identify whether or not when you compared two jobs that had an advance qualification requirement but were slightly different in gender, whether or not those persons treated those jobs differently or not. Then you could try to distinguish between whether or not it was truly a gender preference that was operating or whether or not it was a high education preference.

The gender preference that is identified by the mathematics could be an indirect relationship to some other preference. Basically it is an interpretation question. The mathematics can't really help you, except, I guess, in judgment. The stronger the relationship is, the more striking the difference between the treatment of the male and female questionnaires, the more most people would, in judgment, conclude that it really was a gender preference operating and not something correlated with that.

You talk about a gender preference as opposed to a gender bias. We use the terms "gender preference" and "gender bias" fairly interchangeably in the work because of this concern that someone might get labelled as having a bias when in fact it is potentially related to a preference for education or blue collar, a secondary

relationship. We caution someone that we should call it a preference.

For the two raters that were identified there, the relationship in the data was so strong that I would have a hard time believing that it wasn't gender that was driving the distinction between the way they treated the questionnaires.

426. The IRR Sub-Committee produced its own report concerning the IRR testing performed by Shillington. Their report was released on July 15, 1988, about two weeks before the Tristat Report was officially released. The Sub-Committee's report of July 15 differs from the Tristat Report of July 31, 1988 by referring to 103 "problematic" questionnaires "requiring" re-evaluation. On the other hand, the Tristat Report identified these questionnaires as "unusual" and suggested that they "should be reviewed" not re-evaluated. Shillington testified the IRR Sub-Committee's report used stronger language than he used and, in his opinion, the identified questionnaires should be "looked at, nothing more."

427. Shillington attended the JUMI Committee meeting of July 15, 1988, when the IRR Sub-Committee report was tabled. The tabled report written by a management representative indicated 103 of approximately 500 benchmarks had been "influenced" requiring further examination and possibly re-evaluation and sore-thumbing.

428. Shillington testified the use of the word "influenced" in that context did not reflect what was agreed upon in the IRR Sub-Committee. He indicated his July 31, 1988 report was his best recollection of the opinions formed by the IRR Sub-Committee. Shillington testified he was of

the view the identification of the 103 questionnaires as having been influenced was not supported by the research he had done.

429. Willis also testified about this aspect of the Tristat Report (Exhibit HR-39) and the IRR Sub-Committee Report (Exhibit HR-11B, Tab 26B). Willis did not agree with the section of the IRR Sub-Committee report which dealt with "influential raters" and he stated the sub-committee appeared to overlook the fact that Willis considers it necessary and desirable that evaluation members be permitted to make adjustments in their evaluations at consensus time based on factual information. The fact there is a shift from the majority evaluators towards a minority evaluator in a number of cases is not by itself evidence of a problem. Willis testified in Volume 38, at p. 4803, lines 7 to 20:

There were, if I recall, a couple of raters on the evaluation committee who did have an influence not because they were biased but because they were bright analytical people that others respected. Usually when they had a statement about an evaluation or were asked to provide information relative to the facts of that rating, they generally had very sound reasons, and these reasons were respected. So there were occasions where other members of the Master Evaluation Committee did respond to them.

I don't consider that a limitation. I think that was one of the steps that was built into the process.

430. Eventually, Willis did follow through with a re-evaluation of the 103 questionnaires identified by the sub-committee. Willis and his consultants had been asked by the IRR Sub-Committee to question the assumption the 103 questionnaires presented a problem. One of Willis' consultants, Jay Wisner, did the re-evaluations and prepared a report for the JUMI Committee. His analysis is contained in a report to the JUMI Committee entitled Analysis and Conclusions Concerning the Master Evaluation Committee's Work and dated July 1988 (Exhibit R-22). Willis testified he reviewed each of Wisner's evaluations and made some minor changes in the report. Willis testified it was "our" conclusion that a systematic review of further evaluations was not warranted, nor was a reconvening of the MEC necessary. Willis felt the evaluations were appropriate, and he was very comfortable with the overall results given the reasonable random disparity among the evaluations. Moreover, he felt at this point the JUMI Study should proceed as planned.

431. Willis' conclusions are contained in a report dated July, 1988 to the JUMI Committee. The conclusions concerning the re-examination of the IRR analysis is reproduced as follows:

After careful and intensive consideration of the questions raised by the IRR committee's analysis of the MEC's evaluations, we find that the principal recommendation of that report, that the MEC should be convened to re-examine a large number of evaluations, is not supported.

We have re-examined the evaluations which the IRR analysis indicated were "unusual." We did not find evidence that any raters exercised "undue influence" over the group consensus evaluation. In our opinion, the great majority of the evaluations listed by the IRR committee are the product of accurate and

consistent application of the evaluation plan by the MEC, and should not be changed.

For those few positions where we recommend re-evaluation, we found no pattern of influence by a minority resulting in evaluations with which we disagree; in some cases, we recommend movement further from the middle of the initial individual evaluations. We believe that the eventual re-examination by the MEC of the ten evaluations where we suggest some revision need not delay the convening of the five sub-committees. We recommend that these reviews be combined with reviews of benchmarks sought by one of the five sub-committees.

We have no significant concerns regarding the MEC's understanding and application of the evaluation plan. The MEC's pattern of application of the evaluation plan to positions (their "discipline") differs in some respects from the pattern which the consultants would use. However, given the manner in which the MEC membership was determined, their discipline constitutes a more accurate reflection of the values of positions as commonly understood within the Government of Canada than the consultant could determine from an outside point of view. This kind of adaptation of the plan to the climate and conditions of an organization by an evaluation committee is expected and proper. We would be concerned if there were evidence of inconsistent application of the evaluation factors within or across position families. We did not encounter any evidence of such inconsistency. We believe that the framework of benchmark evaluations and the selection of principal benchmarks by the MEC provides a sound basis for the evaluation of the remaining positions by the five sub-committees. We have found no significant cause for concern and support the progression of the study as scheduled.

(Exhibit R-22, p. 8)

432. Of the 103 MEC questionnaires re-evaluated by Wisner and reviewed by Willis, ten were evaluated differently by the consultants and of these ten, only three were significantly different. Of the three, one was more significant than the others. It was Willis' judgment if the MEC was going to be reconvened, it would only be to review that one questionnaire which was more significant than the others.

433. By September of 1988, the management side of the JUMI Committee were still not satisfied with the manner in which the Willis Plan had been applied by the MEC and continued to express concerns about the MEC's work.

At the September 15, 1988, JUMI meeting, the management side indicated a further analysis should be carried out on problematic benchmarks referred

to in Willis' report of July, 1988 on MEC evaluations. The management side identified 100 benchmarks with problems and forwarded 46 of these 100 to Willis with a list of questions, observations and anomalies. In response to management's request, Willis & Associates conducted an independent review of these questionnaires and attempted to do a "fresh" evaluation, without regard to the MEC's prior evaluations but consistent with the general evaluation discipline established by the MEC.

434. A report of this work was submitted to the JUMI Committee in September, 1988 (Exhibit R-28). This analysis was done by the Willis consultant, Wisner. Willis & Associates agreed with management on a number of their challenges but, in the end, did not identify the existence of a gender pattern. As to the discipline adopted by Wisner in his independent evaluation of the 46 questionnaires, Willis said in Volume 56, at p. 6936, lines 3 to 11:

> I think he was familiar enough with the Master Evaluation Committee's evaluations at this point. We had had discussions as to where they were conservative, where they were a little bit liberal, so that he was able to track, but fairly independently. I would say, though, that while it is not critical, it would appear that he was just a hair more liberal on the average than the Master Evaluation Committee.

435. In the final analysis, Willis and Wisner identified one evaluation out of the 46 which they considered was misunderstood by the MEC. Willis' report provided explicit answers to each of the questions raised by the management side. In their judgment, the additional analysis supported their conclusion the MEC had done a fully satisfactory job in applying the evaluation system to a broad range of positions. The report states:

> We believe that a sound basis has been provided for evaluation of the remaining 3900 positions and that, at this stage, there is no logical reason to expect less than a high quality, defensible result from the study.
> (Exhibit R-28, p. 4 of the addendum)

436. In this report, Willis also provides general observations as to how differences can occur between the MEC and the consultants. The report

states they can occur in three different ways and anyone of these three ways could be caused by systematic bias on the part of evaluators. He identifies the three different ways as follows:

> 3. Differences in evaluations of the same positions between the MEC and the consultants could occur in three different ways:
>
> Misreading of the questionnaire. This could result if parts of the questionnaire were overlooked or not given appropriate consideration.
>
> Different interpretations of the facts given. The consultants may draw interpretations from a more

> extensive experience in evaluating other jobs having similar functional responsibilities. On the other hand, evaluation committee members may have a better understanding than the consultants of the culture within the governmental organization resulting in slightly different job perspectives.
>
> Misuse or misunderstanding of the evaluation system. This is expected only during the learning stages of the evaluation effort.
>
> Any one of these three ways could be caused by systematic bias on the part of evaluators.
> (Exhibit R-28, pp. 1-2 of the addendum)

437. By late November, 1988, the management side of the JUMI Committee were still dissatisfied with the Wisner/Willis analysis of the MEC evaluations. Ouimet forwarded a four page letter to Willis detailing her concerns (Exhibit HR-19). Willis responded to her letter on December 5, 1988, in a six page letter in which he attempted to deal with those concerns. Part of that letter is reproduced as he attempts to persuade Ouimet that some variance between evaluators will occur and the reasons for this variance. He says:

> Evaluation Tolerance
>
> As I indicated in the Addendum to the Responses to the Management Side of the Joint Union/Management Initiative on Equal Pay for Work of Equal Value, it is expected that some variance in

interpretation of position information provided to evaluators will occur. A tolerance of plus or minus 10 percent in random evaluation variance is acceptable between two teams evaluating the same positions, given complete and accurate factual information.

As a practical matter, analysis and assessment of evaluation reliability requires making judgments considering a number of variable factors, such as:

-Completeness, factual content and definitiveness of the information used. Lower quality of information normally results in wider random bias.

-The nature of the job. Is it unusual or complex, or one the evaluators should be reasonably capable of understanding (e.g. research scientist or cleaner)? To evaluate a position properly, the evaluators must be able to understand its content.

-How far removed is it in organizational level from the experience or knowledge of the evaluators? This is similar to the previous factor in that evaluators may have trouble conceptualizing a job that is several organizational levels above their own experience.

-Do evaluation variances depict a pattern? Does there appear to be a systematic bias, or is it a random variance? Systematic bias is much more significant than variance that is simply difference in interpretation or understanding of the job's requirements.

-If the comparison evaluations are by a consultant, could the deviation result from difference in understanding of the culture or value systems within the organization, resulting in a slightly different job perspective?

In essence a value judgment must be made as to the extent of allowable variance in scores and whether or not a problem exists. An assessment of this nature does not lend itself to "precise and quantitative terms".

Of the fourteen MEC evaluations assessed as differing by more than 10 percent compared to consultant evaluations it was my considered

judgment that one, MEC
#428 Head Display Preparation Section, was not properly understood by MEC and should be submitted to that committee for questions to be asked, and re-evaluated.
(Exhibit R-35, pp. 3-4)

438. In the final analysis, the management side of the MEC did not whole-heartedly support the MEC benchmark evaluations. Although they were prepared to continue the study, their intention was to conduct further reviews of the benchmark evaluations, this further review was not addressed by the Employer in the presentation of their case.

439. The Tribunal heard limited testimony from evaluators on the subject of MEC challenges. Pauline Latour, one the evaluators who testified before us on the issue of committee challenges to benchmarks, viewed the MEC challenges as a small issue. Only one particular benchmark caused her committee, (Committee #5), difficulty. It was this committee's view that the position was rated higher by the MEC than what it ought to have been. (Volume 171, pp. 21641-43).

(ii).IRR Testing in the Multiple Evaluation Committees

440. Shillington also conducted IRR testing on the remaining five and nine evaluation committees using the same methodology he used to identify outliers on the MEC. Willis was provided with two written reports on the IRR testing of the five and nine evaluation committees. The first disclosure made to Willis occurred in May of 1988 and was primarily based on the original five evaluation committees. The second disclosure occurred in July of 1989 and was based on the expanded nine evaluation committees.

441. The IRR Sub-Committee reported to the JUMI Committee, at its meeting of August 25, 1989, that an analysis of individual ratings to the end of July, 1989, revealed 11 outliers, six female evaluators from the staff side, three female evaluators from the management side and two male evaluators from the management side. Seven of these outliers expressed an

apparent preference for male positions and four expressed an apparent preference for female positions.

442. The sub-committee further reported seven of the outliers had been previously identified in the earlier disclosure. However, Willis, in his testimony, was able to recall eight outliers who had been previously identified. The identification of the previously identified outliers was

reported in the second disclosure by the IRR Sub-Committee in order to confirm Willis' opinion as to the ineffectiveness of his intervention/counselling following the first disclosure.

443. The JUMI Committee decided the names of the outliers would only be revealed to Willis and the Chief of the EPSS. It was Willis' understanding that the JUMI Committee's decision to deal with the question of outliers in this confidential manner, was done to protect the individuals concerned. The JUMI Committee had made an earlier decision they were not going to remove any evaluators from the committees and it would not be productive to release their names at this point.

444. Shillington prepared exhibits identifying what he referred to as "an underlying attitudinal dimension of these outliers". He was unable to explain why these differences occurred or what they were. Exhibits HR-117 and HR-133 indicate the male and female preferences crossed union/management lines and female/male lines. With respect to the cross over of male/female lines, some female evaluators displayed a male preference; however, no male evaluators displayed a female preference.

(iii). Inter-Committee Reliability Testing

445. Willis testified inter-committee reliability (ICR) testing is designed to determine whether evaluations from a series of committees are related. As explained by Willis, this testing looks at consistency between committees and identifies where committees need to be retrained. Willis testified ICR testing is not designed to identify any form of bias. In the JUMI Study, it was intended instead, as a means for assessing whether or not the evaluation committees were adapting successfully to the discipline of the MEC.

(iv). ICR Testing in the Multiple Evaluation Committees

446. The process generally involved taking a series of questionnaires and submitting the questionnaire to each committee. Each of the evaluation committees performed an evaluation on the same questionnaire and the consultant then attempted to identify the extent to which different committees rated the same job similarly or rated the job differently. According to Willis the first ICR testing began early in 1989 and included 26 tests altogether. The ICR testing continued until July of 1989.

447. The JUMI Committee established an ICR testing sub-committee (the "ICR Sub-Committee") to establish policy and oversee procedures for the testing. The ICR Sub-Committee consisted of three management representatives, two staff representatives, Willis, one of his consultants and two Commission representatives. The purpose of the IRR Sub-Committee

as stated in the IRR Sub-Committee report of March 3, 1989 is listed as follows:

-examine the results of the tests administered to the evaluation committees in relation to the baseline provided by the consultants,

-examine the baseline score provided by the consultants,

-determine the significant differences in the consensus ratings of the committees in relation to the benchmarks and the baseline,

-formulate if needed, recommendations for training, re-training by the consultant and/or other courses of action for JUMI considerations, and

-identify procedural/process problems and potential for improvement including the revisions to the formulation of rationales.

448. The IRR Sub-Committee requested the Commission conduct the actual testing. The Commission determined the timing of the tests, distributed the questionnaires and explained the process to the committees. The JUMI Committee asked Willis to evaluate the test questionnaires and provide a "baseline score" for each of the test jobs.

449. The baseline score was the independent evaluation of the test questionnaires by the consultants. In each case, Willis had two consultants review the questionnaire and arrive at their own independent evaluation, which was then compared with the test evaluations done by the five or nine committees. The purpose of the comparison between the baseline score and the committee score was to identify any deviation between, first, the individual committees and, second, the consensus of the multiple evaluation committees compared to the consultants' evaluations, thus identifying areas where the multiple evaluation committees needed to be retrained because of difficulty in interpreting the evaluation factors.

450. Willis used rationales in the ICR testing to analyze differences between the baseline scores and the committee scores. His use of the committee rationales in the ICR testing was for a different purpose than the use of rationales generally in the evaluation committees. In the ICR testing Willis explained why rationales were to be used in this exercise, as distinct from his reasons for not wanting them to be used in evaluations

by the committees, in which case he wanted the members to focus on the questionnaire itself.

451. The consultant baseline score was compared with each committee's consensus score and also to the overall consensus of the five, and later nine, evaluation committees.

452. Willis had minor input into the procedure that was adopted by the ICR Sub-Committee and he was opposed to their approach. In other studies, Willis always provided a list of questionnaires to his clients, and then

introduced the questionnaire into the committee's portfolio of questionnaires in such a way the committees did not know which questionnaires were part of the test. In the case of the JUMI Study, time was set aside and the questionnaires were distributed to the evaluators who then became aware of the testing. The Commission randomly selected the questionnaires and approached the Willis consultants about an hour before the test to give them an option as to which questionnaires should be used for testing. The consultants were not given the opportunity of selecting questionnaires that were more complete. As a result, Willis testified there was frustration on the part of the evaluators, as well as varying levels of conscientiousness in completing the tests.

453. The procedure for ICR testing as conducted by the Commission was very strict at first. It was announced there was going to be a test. The Commission was on hand to oversee the test and had an observer in each room. The questionnaires were distributed and the evaluators were informed they could not leave the room during the actual period of the test.
454. In other studies, if the committees needed more information on the test questionnaires, Willis arranged to have individuals waiting at telephones to answer any questions. No time was permitted for this in the actual ICR testing. Consequently, each committee was allowed to make its own assumptions and fill in any gaps in the information. Committees were required to write down their assumptions but the problem was that each committee made different assumptions. Willis testified because committees were making different assumptions, variance occurred in the evaluations. For these reasons Willis was not comfortable with the results of the ICR testing.

455. Willis found the committees did not take the ICR tests as seriously as they did their actual evaluations. He observed a considerable amount of resentment on the part of the evaluators and this increased over time. Moreover, the committees had to stop their regular evaluations to go

through the testing exercise. The committees were being pressured by Willis to keep moving but, at times, were subjected to two tests a week. As Willis stated in Volume 58, at p. 7166, lines 13 - 16:

> They resented it every step of the way and some of them quite frequently took the testing with somewhat less than a serious approach.

456. On February 6, 1989, Willis produced a report on the first nine tests conducted between November 7, 1988 and January 5, 1989. This report examined the variation among the original five evaluation committees and essentially concluded they had learned the Willis system and were evaluating positions in line with the MEC discipline when "they feel comfortable with that discipline."

457. The ICR Sub-Committee report of March 3, 1989 was based on an analysis of the first 11 tests conducted. The report noted: (i) that the consultants needed to go through a revision of the initial training program with the committees and to address problems that were identified; (ii) there was some concern with respect to evidence of cross-family job comparisons and the job evaluation process ought to be amended to provide

for these comparisons; and (iii) the rationales needed more attention and a revised job evaluation process ought to be developed.

458. Willis was asked to describe the amount of variance between the consultant scores and the consensus of the five committees on the first 11 tests. He responded in Volume 58, at p. 7227, lines 1 to 9, as follows:

> A. Considering the various handicaps and expressions of frustration and concern that we heard, I think that they did very well. I felt very positive, particularly after discussing with each committee what their differences were, why they had selected the assumptions they had. While I did agree that the additional re-training was desirable, I felt very positive about how well they were doing.

459. The ICR Sub-Committee attached to their report a description of an Improved Evaluation Process which it recommended for adoption by the committees. The revised process provided for comparisons of benchmarks outside of the job family. It asked evaluators to first reference benchmarks in relation to their factor ratings before independent ratings were passed to the committee chair for posting.

460. Willis testified the committees tried the improved job evaluation process and found it was not really practical and actually required more time than the original process. The evaluation committees resisted the change so it was finally dropped.

461. Willis also produced a report on the ICR testing. He described the sub-committee's report as similar to his own report for the most part, but did not completely concur with all of the sub-committee's findings.

462. The essential purpose of the ICR testing was to identify whether or not committees understood and were applying the Willis Plan in accordance with the MEC discipline. This testing also gave the consultants an opportunity to examine whether or not the committee's reasons for their ratings were suggestive of gender bias.

463. Willis did a careful review, factor by factor, of what each committee did, why they did it, and how the consensus was reached for each factor and for the total. When one committee's score differed from the other committees' scores, Willis explored the reasons for the difference. If those reasons suggested, in any way, they were influenced by a particular gender or by a particular kind of job, it was information that would be available to the consultant for follow up action.

464. Within this testing framework, Willis was asked whether he found any evidence of gender preference in the work of the committees during the first series of ICR tests. His response is given in Volume 58, at p. 7227, line 10 to p. 7228, line 19:

> Q. At this point -- and it looks like we are in the late winter or the early spring of 1989, when only half of the tests had been performed -- did you have any evidence from these tests, or

> otherwise, that there might be a problem of gender preference exhibited by the evaluators?
>
> A. Unrelated to the ICR testing, in the early part of the year I began to express some concerns, not related directly to the actual evaluations themselves, but I had some concerns regarding some circumstantial things that had been happening. I became increasingly uncomfortable with how the committees were working with the confrontations between the staff and the management side, and some of the circumstantial things that I had observed happening.

In stressing with my consultants working with the groups, and doing our own analysis of how committees were actually evaluating, and how occupational groups were coming out among and between the committees, I could identify nothing specific that would suggest there was a gender bias that was developing.

Nevertheless, I had strong mixed emotions because I knew some things that were happening, some attitudes that were apparent that were giving me a great deal of concern.

So at this point in the study I had some problems with my own level of comfort. I discussed these problems individually and with the members of the mini JUMI and collectively with them as a group. I felt that I was going to have to take some sort of an analysis, a more in depth analysis of results if I was going to be able to support the outcome of the study.

465. Willis did not think it possible to identify gender bias simply by looking at the results of the ICR testing. If there is gender bias, Willis finds evaluators usually tend to talk about their conclusions or opinions rather than about the facts of the questionnaire. He indicated to his consultants to watch very carefully for this sort of behaviour but he does not think a consultant can decide whether or not there is bias just by looking at a score or on a job by job basis. He testified on an individual job evaluation, a consultant has to look at the reasons why the committees selected what they did, what was stated in the rationale and then quiz the evaluators personally as to what were the reasons for the differences. In Willis' opinion these tests did not provide any conclusive evidence of gender bias and the information obtained from these tests should be discounted because the committees did not take the testing seriously.

466. In late May or early June of 1989, Willis recommended to the ICR Sub-Committee the testing be discontinued because it was becoming very clear to him the evaluation committees were becoming more and more frustrated with this procedure. He also concluded the tests were beyond any point of usefulness. Willis understood it was at the insistence of the Treasury Board representative on the sub-committee, that the testing continue. The sub-committee did not accept his recommendation and continued with the testing into July of 1989. It was Willis' opinion the reaction he observed by the evaluation committees to these tests might have an affect on the reliability of the results. (Volume 59, p. 7291).

467. Although the testing continued, Willis did not perform any additional formal analysis on the results. He reviewed the remaining tests submitted to him by the sub-committee and continued to meet individually with committees.

468. A draft final report of the 26 ICR tests was prepared by a management representative, Michel Papineau, on the ICR Sub-Committee. This report is dated October 26, 1989. Willis had no input into this draft. The conclusion reads as follows:

> The ICR test results tend to support the gender preferences found in the IRR report and in the Consultant's study of a sample of 220 questionnaires already evaluated by the committees. The differences are such that there is little doubt as to whether or not these are due to systematic or random biases. The proportion of these discrepancies are significant enough to exceed the degree of tolerance expressed by the consultant. Thus, it is strongly recommended that further investigations be conducted prior to reaching any conclusion based on the evaluation results.
>
> (Exhibit HR-90, p. 4)

469. Papineau concluded there was evidence of gender bias in the evaluations but it was Willis' judgment, the analysis of the ICR testing should be discounted for two reasons. He stated the 26 evaluations were too small a number from which to draw any firm conclusions, and secondly, the committees were not taking these tests as seriously as the actual evaluations and were rushing through them as quickly as they could without much discussion. It is Willis' opinion the tests were "not valid for any particular use after about the first 10 or 12 tests." (Volume 59, p. 7297).

470. As to the assertion in the report that further studies should be undertaken, Willis testified he had already decided on the basis of the Wisner 222 a further study needed to be undertaken and this draft ICR report did not add to his conviction.

471. According to Willis, the Wisner 222 was not related to the ICR testing at all. He testified he would have asked for the Wisner 222 whether or not the JUMI Committee had agreed to conduct the ICR testing. Willis saw it as a totally separate issue.

472. Elizabeth Millar, a union member of the ICR Sub-Committee, employed by the Alliance as Head of Classification and Equal Pay Section, testified she was under the impression the ICR testing was being taken very seriously by the committees. She testified one of the problems of the ICR

Sub-Committee was in getting timely feedback to the committees. Millar said she did not think the ICR Sub-Committee functioned in an effective manner after May, 1989. She stated the management representatives on the ICR Sub-Committee appeared to adopt a different agenda from the rest of the committee. These representatives wanted an increase rather than a decrease in the schedule of testing to the end of the evaluation process.

473. By memorandum dated November 10, 1989, Millar responded to the draft report prepared by Papineau. Essentially, she found the draft unacceptable to the Alliance as it did not reflect the discussions and deliberations which took place within the ICR Sub-Committee. The analysis contained in the report did not reflect the committee's findings and the conclusion contained in the report had never been discussed by the ICR Sub-Committee. In his testimony, Willis agreed with Millar that the concluding statement contained in the draft report was perhaps overstated. It implied the comparison left little doubt as to the existence of gender bias. (Volume 59, p. 7304).

474. In Papineau's memorandum, which is attached to the minutes, he indicates his intention to table the report at the next JUMI Committee meeting which was held on October 31, 1989 (Exhibit R-44). However, the final report was not tabled at this time, since the report had only been distributed one week prior.

475. Millar testified in the spring of 1989, she observed a change in attitude by the management representative on the ICR Sub-Committee toward the consultants. She said the Employer's attitude before May of 1989 was more accepting of and in tune with the consultant's view so that the sub-committee was able to reach agreement in problem areas. It was agreed the evaluation committees had trouble understanding the Willis Plan and needed further help in training. She testified after May, 1989, the Employer representatives became very critical of the consultants and the ICR Sub-Committee meetings became extremely difficult. She recalled one particular meeting in which Scott Gruber, a Willis consultant, reported on one of the tests that had been done. Gruber had met with all committees to discuss the results and found overall the work was going well. The Treasury Board representatives took issue with Gruber's report. According to Millar, one Employer representative commented to the effect the committee ought to have expected something better from the consultants.

476. Millar referred to another incident in Volume 185, at p. 23775, line 17 to p. 23776, line 11 as follows:

> At one meeting in which Mr. Owen was the consultant, two Treasury Board representatives turned up with reports that we hadn't known were in the preparation which had calculated the difference between each committee score and the base line score and had used these calculations to indicate whether or not a problem had existed.
>
> Mr. Owen, who I have described as unfailingly polite and a kind individual, as well as very competent, became extremely agitated. He threw his pencil across the desk and accused both the Treasury Board representatives of neither understanding job evaluation or the Willis Plan. Mr. Willis reported to me later on that he had worked with Fred Owen a long time and he had never seen him so angry. Needless to say, these reports, the uncommissioned reports, were never accepted by the subcommittee and were never tendered further.

Mr. Owen was not questioned about this incident.

477. Willis testified the ICR testing fell short of his expectations. He said for future ICR testing, he would arrange to do it covertly so the evaluators would not know they were being tested. He did comment concerning the ICR testing results as follows in Volume 59 at p. 7352, lines 2 - 7:

> But the bottom line is, apparently, in spite of lack of management support, in spite of some variances in the quality of information and in spite of some attitudinal problems, the result was within satisfactory limits.

478. Willis was also asked whether there was an indication in the first 11 ICR tests that the committees valued higher along gender lines. He testified it was his assessment there did not appear to be a gender preference. Any differences in interpretation between the consultants and the committees on the "problem-solving" factor in the Willis Plan was due more to a lack of clear understanding of how to use the evaluation system than anything else.

(v). Wisner 222 Re-Evaluations

479. Willis testified it was clear to him there were agendas both on the staff side and on the management side affecting the way evaluators worked together. He observed attitude problems on the part of some of the

evaluators. As the study proceeded, Willis became concerned that he could not defend the results without doing further analysis. Willis' discomfort did not result from what he was able to observe, in terms of actual gender bias in the evaluations. It centred primarily around what he viewed as an attitude problem on the part of the evaluators. Because this was a large and important study, he wanted to be sure there was no subtle bias creeping into the process.

480. Willis made a recommendation to the JUMI Committee to conduct a "snapshot" assessment on the validity of the evaluations, with the intention that if his preliminary analysis revealed the possibility of problems, he would subsequently do a more in depth analysis. When Willis made his proposal to the JUMI Committee in the spring of 1989, he did not advise that he anticipated adopting a two-tiered approach if a problem was encountered in the first small study undertaken. His recommendation to the JUMI Committee was made about the time the first 11 ICR tests had been completed. At this point, the committees had evaluated approximately 2000 questionnaires, and Willis wanted to examine 10 per cent of these completed evaluations using one of his consultants to independently evaluate a sample.

481. Willis testified the only way he knows of determining whether gender bias is present in an evaluator's evaluation is to look for patterns of preference for one gender or the other. In his opinion, the only possible way of identifying gender bias would have been to have an impartial third party, such as one of his consultants, re-evaluate selected questionnaires, then to compare the results between the committees and the

consultants. Willis usually solicits the assistance of a statistician to perform this comparison. Willis refers to differences between the committees and the consultants as disparities.

482. During the course of the hearing, there were questions about whether consultants should or should not be considered the baseline for comparison. Willis pointed out the JUMI Committee had agreed to use the consultants as the baseline in the ICR studies, and that agreement was expressed in writing by the JUMI Committee.

483. Willis believes his consultants who were involved in the JUMI Study were unbiased and testified to this in Volume 208, at p. 26934, lines 10 to 16, as follows:

We understand the system. I think it would not be appropriate to say that all consultants are necessarily unbiased. However, our experience, our background, our intent, our own philosophy, has always been not to favour one side or the other, but to walk in the middle road, if you will.

484. Willis testified the disparities form the basis for identifying whether or not there is a gender based pattern. In this context, he said "bias" simply means if there is a pattern of different treatment for male-dominated jobs versus female-dominated jobs, then the different treatment would result in some degree of gender bias.

485. The positions included in the Wisner 222 were selected randomly from a list of all the evaluations provided by the EPSS. The sample taken included at least 10 per cent of the total number of positions evaluated by the nine evaluation committees at the time of the Wisner 222. The sample included the full range of evaluation levels and the variety of types of work seen by the nine committees.

486. Wisner did not testify at this hearing. His study was explained by Willis who described the method used by Wisner in his analysis. First, Wisner read the position questionnaire and any reviewer notes. He then determined whether a similar position was included among benchmark positions evaluated by the MEC. When there was a similar position, Wisner reviewed the benchmark questionnaire to confirm his impression and adopted the MEC benchmark evaluation as the consultant evaluation. When there was no similar set of duties among the MEC benchmarks, Wisner proceeded to do an independent evaluation of the position, supported by reference to appropriate benchmarks. Many of the positions included in the sample were found to require this step. After determining an evaluation, Wisner reviewed the committee evaluation for that position. He paid particular attention to the committee's use of benchmarks and the facts they used to support their evaluation. Wisner then adjusted his evaluation as appropriate in view of the committee's rationale and benchmark references.

487. When Wisner found differences between his final evaluation and the committee's evaluation, he wrote a brief rationale in support of his position.

488. Wisner then proceeded to do a special analysis on the results. This analysis was initiated in order to assess the quality of the position evaluations by the nine evaluation committees. As stated in his report of

July, 1989, the considerations he included in determining the "quality" of the evaluations were:

1. Proper use of the Willis evaluation system in accordance with the Guide to Position Measurement and the training and technical advisories issued by the consultant.

2. Consistency of the evaluations by the nine committees with the benchmark evaluations and evaluation discipline established by the Master Evaluation Committee.

3. Absence of any systematic bias in the evaluations by the nine committees.

(Exhibit PSAC-4, p. 1)

489. Wisner's analysis also included statistical testing. According to Willis, Wisner is a statistician. His findings on the first consideration regarding the proper use of the Willis evaluation system, was that there was no evidence found, with two possible exceptions, of any consistent misinterpretation or misapplication of the evaluation factors and dimensions. As to the two exceptions, he noted because the number of positions sampled were so small that it was impossible to draw any firm conclusion about these.

490. Regarding the overall consistency of evaluations by the nine committees with the MEC evaluation discipline, he found that the committee and the consultant had an exact match in 70 of the 222 positions, and that an additional 34 positions showed differences of +/- 2.5 per cent, so that almost 47 per cent of the positions in the sample had approximately the same overall evaluation. He concluded these differences indicated fair consistency of evaluation between the nine committees and the MEC benchmarks. Since Wisner found more than half of the positions differed by more than 2.5 per cent, he recommended that further analysis of the differences was warranted.

491. As to the third consideration, in analyzing the differences between the consultant evaluations and the committee evaluations, Wisner found for the female-dominated positions, 35 were under-evaluated compared to the consultant, 40 were over-evaluated and 43 had no difference; and for the male-dominated positions, there were 55 under-evaluated, 22 over-evaluated and 27 with no difference. His report states at p. 5:

This indicates that female dominated positions were over evaluated somewhat more often than the total sample, and male dominated

positions were under evaluated somewhat more often than the total sample.

492. And his conclusions at p. 8 reads:

> The findings of the analysis described above suggest that the consistency of the evaluations by the nine committees with the MEC benchmarks is less than would be desirable, and that there may be some gender-related bias in the evaluation results. It is the consultant's opinion that these findings indicate that a wider review of the evaluations by the nine committees would be proper. Such a review would serve to confirm or refute the apparent problems found in the sample of positions examined in this study. [emphasis added]

493. Wisner, however, advises caution in dealing with his report. The statistical analysis between gender dominance and evaluation differences between the committee and himself are based on a comparatively small number of positions and his findings "between the two variables does not mean that there have been deliberate or unconscious sex bias in the evaluations." He goes on to say there are a number of other possible explanations for the differences. He refers, for example, to the tendency in the positions in the male-dominated classifications to have more complex duties and responsibilities than the majority of positions in the female-dominated classifications. He suggests the observed pattern of evaluation differences could occur if the committees tended to under evaluate more complex positions in relation to the MEC discipline as viewed by the consultant.

494. Willis' covering letter of July 17, 1989, addressed to the co-chairs of the JUMI Committee, which accompanied Wisner's report, states in the third paragraph:

> Our findings indicate the existence of some systematic divergence from MEC evaluations. Statistically, however, the size of the sample reviewed, 222 evaluations, was insufficient to permit specific conclusions as to the degree of the problem.

(Exhibit PSAC-4)

495. Willis was asked in Volume 58 to clarify exactly what it was he was trying to state in this letter. He responded in Volume 58, at p. 7249, lines 1 - 5, as follows:

A. The results of our analysis appeared to suggest that there is some pattern of deviation from the Master Evaluation Committee's evaluations. It could be interpreted as a gender bias.

496. At the completion of the Wisner 222 there were about 1,000 evaluations remaining. Since the nine evaluation committees had just started their work, Willis felt it was critical that a more extensive analysis be done as soon as possible to correct a potential problem. He recommended to the JUMI Committee an additional analysis be undertaken without delay.

497. In his testimony, Willis referred to the following table contained on p. 4 of Wisner's Report to explain why he wanted a further study and his concern about possible gender bias:

Table 1
Per Cent Differences

Group	<-15%	-14.99 to -10.0	-9.99 to -5.00	-4.99 to -2.50	-2.49 to -0.01	0	0.01 to 2.49	2.50 to 4.99	5.00 to 9.99	10.00 to 14.99	>15
Female	6	8	7	5	9	43	9	10	4	9	8
Male	8	15	13	9	10	27	6	4	4	4	4
Total	14	23	20	14	19	70	15	14	8	13	12

498. Willis testified the above table breaks down the total group of the 222 evaluations. In the first line which reads "Female", highlighted under 0, the 43 indicates Wisner and the MEC agreed on 43 evaluations. To the right of 43 is the number of MEC evaluations above Wisner and these total 40, and to the left of 43 is the number of MEC evaluations below Wisner and these total 35. On the "Male" side, highlighted under 0, the number in the chart indicates Wisner and the MEC agreed on 27 evaluations. The right hand columns indicate that the MEC rated 22 evaluations higher than Wisner, and the left hand columns indicate that the MEC rated 55 evaluations lower than Wisner. This suggested to Willis the beginning of a pattern because there are approximately twice as many (55) male-dominated evaluations rated lower than the number of evaluations which agreed with the consultant (27) and the number (22) which were over-evaluated compared to the high number of male-dominated jobs within the total male-dominated

occupational groups which were evaluated lower by the committees than the consultant's evaluation.

499. This aspect of the Wisner 222 concerned Willis. Another concern with the report was that it showed one female-dominated occupational group (ST) in which the numbers indicated a comparatively large degree of over-evaluation. This to Willis was some evidence, however slight, of gender bias.

500. Willis stated the Wisner 222 was very limited. It was not intended for a basis on which to make a determinative judgment as to whether or not true gender bias existed and to what extent. He testified it contained enough evidence to justify a further look before he could feel comfortable in defending the results.

501. Following the release of the Wisner 222, the unions sent a letter to Durber, expressing their concerns. This letter is dated September 27, 1989. The letter, which is written by Christine Manseau, the union co-chair, indicates the unions did not agree the Wisner 222 Report supported

the contention there was gender bias in the evaluations. Paragraph 2 of the letter reads as follows:

> Our analysis shows that, on average, there is remarkably little difference between the evaluation scores of the consultant and the committees. Of the 118 female positions in the sample, the average consultant score is 182 and the average committee score is 181. Of the 104 male positions, the average consultant score is 273 and the average committee score is 263, a difference of 3.7%. We do not believe these differences are significant and we note that they are well within the + or - 5% accuracy level for average scores that the parties agreed to in dealing with the issue of sample reduction and overall sample size for the JUMI study.
>
> (Exhibit PSAC-5)

502. According to Kathryn Brookfield of the Institute, after having received the Wisner 222, the unions expressed concerns as to how the data in the report matched with the conclusions. Brookfield testified the union looked at the distribution of the evaluations from female-dominated occupations and did not see evidence of an imbalance in the evaluations and yet the report came to that conclusion. Brookfield testified the unions wanted to sort out in the Wisner 222 why the data and the conclusions did

not agree. Until that question was resolved, the unions did not have sufficient confidence to ask Willis to go ahead and repeat the exercise. Brookfield further stated the unions wanted to meet with the Treasury Board representatives, go through the report, discuss the differences and see if they could come to some understanding about them.

503. There was considerable debate within the JUMI Committee as to whether Willis should undertake further re-evaluations. Willis met privately with members of the Mini-JUMI as well as with the full JUMI Committee to request a more in depth analysis. He never wavered from his position that a further analysis was needed, although, the extreme positions taken by some evaluators seemed to settle down during the course of the summer of 1989 as the committees began to work with new, fresh, and in some cases, reorganized committees. He said in Volume 58, at p. 7285, lines 4 to 8:

> A. Call it a gut feel, I just felt that the importance, the size of the study was such that I wanted a better feeling of confidence that I could, in fact, defend the results.

504. At that time, Walt Saveland, an employee of the Commission, did a "technical examination" of the Wisner 222 analysis. Saveland was a staff person with the Commission in Policy and Research. Durber had asked him to assist in interpreting the Wisner 222 and to pinpoint the problem of bias. The Saveland Report, Exhibit PSAC-6 entitled "Technical Observations and Suggestions on Willis & Associates "Special Analysis of Working committee Results" provided a list of male jobs which the Commission ought to give priority attention because the committees differed from the consultant by 10 per cent or more. In the end, the list contained 25 jobs,

notwithstanding 27 had been identified. Wisner and the committees agreed with an additional 2 jobs which had somehow been missed.

505. The Saveland Report appears to pinpoint the source of apparent gender bias to the male-dominated questionnaires. The balance of the Saveland report, from page 6 onward, uses a number of statistical measurements which, according to the statistical expert, Sunter, are "absolute nonsense". (Volume 105, pp. 12696-97).

506. Paragraph 2 of this report states as follows:

> The most important evidence of apparent gender bias is found among male-dominated jobs. A pivotal role seems to be played by 27 jobs

> in which Committee evaluations where [sic] between 5 and 15% lower than Consultant evaluations. (Evidence of apparent gender bias was also found among the clerical portion of female-dominated jobs.)
> (Exhibit PSAC-6)

507. Saveland's report states "it is this kind of asymmetry in the male-dominated line which indicates apparent gender bias." Saveland explored the effects of asymmetry by expanding the standard for relative agreement from a +/-2.5 per cent to a standard of +/-5 per cent. If the expanded standard is imposed for the category of relative agreement with respect to the female-dominated line, it results in a perfectly symmetrical distribution with a sizable majority of jobs, showing 76 relative agreements. For the male-dominated jobs, 56 are now counted in relative agreement but apparent under-evaluations out number over-evaluations by exactly 3:1 or 36 to 12.

508. The report contains, in the end, technical suggestions. One suggestion was a re-examination of the specific jobs in dispute, to be done by some existing or newly formed review committee, whose members are experienced in job evaluation. The report states this review committee should consider all jobs in a "suspect" category and "this means all existing and additional male-dominated jobs (and possibly all clerical jobs)." The report notes an examination of only selected jobs playing a pivotal role in gender bias runs the risk of losing objectivity. The report makes suggestions about what approach ought to be used when a review committee accepts or rejects a specific committee evaluation. The report also suggests, while the review committee is doing its work, the consultant could be re-evaluating the same jobs. Wisner would be the preferred consultant for job re-evaluation because according to the report, he offers the best assurance of continuity. The report states at p. 24:

> If others do the work for Willis and Associates, then quality-control procedures should be put in place to make sure that new Consultants would have done the previous work in exactly the same way.
> (Exhibit PSAC-6, p. 24)

509. At the October 31, 1989 meeting of the JUMI Committee, Saveland was in attendance. He presented his analysis of the Wisner 222. (His

report was released subsequent to this meeting and bears the date November 10, 1989.) Brookfield testified that Saveland, in his presentation, had

concurred with the unions' position which was there was no evidence in the report of systematic over-evaluation of female positions. Saveland also told the committee, most of the differences between Wisner and the evaluation committees were found with 27 male positions.

510. Durber also attended the October 31, 1989 JUMI Committee meeting. The minutes (Exhibit R-44) state at p. 9 that Durber requested the JUMI Committee to indicate how it would deal with the apparent gender bias referred to in the Wisner 222. Durber offered the Commission's assistance to the JUMI Committee. At that time, the management side of the JUMI Committee was willing to do further reviews of the Willis results. The staff side position, communicated by Manseau, the union co-chair, was that prior to this meeting, the staff side were not in a position to proceed further with the Willis study. Manseau promised to reply to the management side by November 10, 1989, about whether the staff side would proceed and who would represent the staff side in the joint process.

511. Following the JUMI Committee meeting of October 31, 1989, Durber sent a letter dated November 10, 1989, to Manseau. In his letter, Durber notes the Commission's concern is with apparent gender bias and the Commission had drawn no further conclusion at that time, but expected the parties to resolve the question of bias in a way that would satisfy the requirements of the Act. He referred to the fact that Saveland, in his written report makes reference to reviewing the 27 male jobs, and offers a caution that the separate exercise should be done with care to ensure objectivity.

512. In an attempt to understand the Wisner 222 Report, the unions approached their members who had been on the MEC to obtain information which might assist in explaining the differences between evaluations done by Wisner and those done by the committees. Brookfield testified she received information from the CATCA union. A member of CATCA, Rick Smith, was provided with the 27 male questionnaires and assigned by the union side to analyze these questionnaires. The information he provided was reported and filed as Exhibit PIPSC-129. The author of the report did not testify at this hearing. His conclusions are contained on page 2 and 3 of the report which reads as follows:

> In summary, after careful review of the committee results and consultant results I find that the consultant has been consistently higher in ratings for several reasons. Some are outlined above and others are individually pointed out in his rationales. The % differences which I have indicated between Committee and Consultant range from insignificant (in my opinion), 5.4%, to 17.4% which is just at the edge of an acceptable error tolerance. I can find no evidence of bias nor can I say that I

could discount the possibility. The committees and the consultant have provided complete, sound ratings with logical rationales to support them. They are slightly different in all cases but this is to be expected. My own analysis of the positions was often

slightly different than both or leaning toward the committee or the consultant rating.

The process is not an exacting science and the Willis plan does not provide for a wrong or right evaluation of a job. A consensus is the best one can expect and I have no reason not to accept the ratings of the committees as they stand.

513. According to Brookfield, the unions were anxious to meet with the Treasury Board representatives with all the information the unions had gathered, including Smith's report, supra, to determine if the differences between the consultants and the committees could be explained.

514. Ouimet wrote to Manseau, by letter dated November 27, 1989, indicating the management side required a response to its request that Willis & Associates be instructed to do further work. The letter stated management required a response by December 1, 1989 or they would "proceed unilaterally" (Exhibit HR-17, Document 22).

515. The next meeting of the JUMI Committee was scheduled for December 13, 1989. Brookfield testified there was no opportunity for the union side to discuss with management side the report received from Rick Smith of CATCA. It appears from the letter of November 27, 1989, from Ouimet that the management side had embarked upon a review of the 27 questionnaires identified by Saveland of the Commission. The second paragraph of the letter reads:

As requested, we are prepared to exchange comments on the 27 questionnaires identified in Mr. Willis' analysis on December 8; the modalities of a sub-committee will be discussed at the December 13 meeting. Its work however, is independent of the research required by Willis and Associates; this work must proceed immediately and would be concurrent with that of the committee if it is established. Even if the committee finds an explanation for the 27 questionnaires in question, we still require more evaluations to make bias estimates for the various employment groups in the study. At this late date, delays are a luxury we can ill afford. We require a response from you concerning Willis

and Associates further work by December 1, or we will proceed unilaterally.
(Exhibit HR-17, #22)

516. The union side concluded from its reading of this letter, even if a joint process to find explanations for the 27 male questionnaires was undertaken, Treasury Board intended to proceed with Willis' recommendation for a further study with or without the consent of the unions. This became a reality when the union co-chair received Ouimet's letter of December 11, 1995 which reads in part:

We remain firm in the belief that the uncertainty surrounding these evaluations mandates further study. We accept the recommendation by Willis and Associates to undertake further analysis (supported, it would seem, by the CHRC). We have agreed

with your proposal to examine the 27 evaluations cited by the CHRC as relevant to `apparent bias', but you have not responded to our proposal to proceed with further evaluations at the same time. To quote Mr. Durber `...we are anxious that the matter of gender bias be dealt with quickly'. Your responses to our letters leave us no choice but to conclude that you do not want to resolve this issue in the near future. We have decided therefore, to comply with the recommendations expressed by both the Consultant and the CHRC and to proceed as of December 11 at which time the process by which Willis and Associates may undertake further analysis will commence. We will keep you informed of the progress of the study. You may have our assurance as well, that the same methodology unanimously agreed to by JUMI in the first phase will be carefully followed. [emphasis added]
(Exhibit HR-17, #7)

517. Willis testified the decision of the Employer to proceed unilaterally and authorize him to do additional re-evaluations was announced to the staff side without consulting him in advance. When the December 13, 1989 JUMI Committee meeting convened, a statement was read by Manseau. At the request of Manseau, the statement was appended to the minutes after which the unions withdrew and no further business was conducted. The statement made by Manseau is reproduced in full.

STATEMENT BY CHRISTINE MANSEAU
CO-CHAIR OF JUMI
ON BEHALF OF THE PUBLIC SERVICE UNIONS

For some time the unions represented at JUMI have not felt equal partners in this joint undertaking. We had wanted to discuss jointly the conclusions of the CHRC on the findings of Willis and Associates in an informal setting so that perhaps JUMI could arrive at a joint agreement on how to deal with their recommendations. We had suggested the establishment of a sub-committee to review jointly our conclusions on the consultant's evaluations reported in the Willis Special Analysis prior to proceeding with further analysis - we were denied that. We had asked further analysis not proceed unilaterally for we felt it would endanger the joint character of the Study and undermine JUMI's credibility - we were denied that.

In view of Ms. Ouimet's letter of December 11 announcing that Treasury Board has decided to proceed unilaterally with further analysis by Willis & Associates, we feel this Study is no longer joint. We therefore are not willing to participate in any discussions on any outstanding issue at this time.

We request this statement be recorded verbatim in the minutes and that the correspondence exchanged since the last meeting of JUMI be attached to the minutes.

(Exhibit HR-11B, Tab 34)

518. From the August 25, 1989 JUMI Committee meeting when Willis first recommended a further study to the December 13, 1989 JUMI Committee meeting when the union side temporarily withdrew from the study there was considerable tension between the parties. This tension manifested itself even earlier during the work of the IRR and ICR Sub-Committees but it was after the release of the Wisner 222 that the relationship between the management and union sides began to rapidly deteriorate.

519. From August 25, 1989 onward, the union side wanted to move forward with the JUMI Study to conclude the evaluation phase, to determine the methodology for compensation and wage comparisons and if a wage disparity was identified to continue with bilateral and multilateral meetings as required. On the other hand, from the August meeting, the management side felt strongly that an additional study was required and the matter of apparent gender bias could not be dismissed without this study.

520. As the parties became more entrenched in their positions throughout the fall of 1989 the tension escalated. Between November 7,

1989 and December 11, 1989, there were no less than 21 letters introduced into evidence written between the JUMI co-chairs with as many as three letters written by one side on the same day. As Brookfield said in Volume 169, at p. 21296, line 24 to p. 21297, line 9:

> Q. Had you ever had that kind of flurry of paper before in the years that you had been involved in dealing with each other?
>
> A. No. I think HR-17, over, I think we are talking, a six-week period, every issue imaginable about several -- four or five, issues are going on with correspondence, some it [sic] simultaneous, and I think that speaks rather directly to the fact that people were having a lot of difficulty communicating with each other, that there was this flurry of correspondence.

521. Since the unions refused to go along with a further analysis, Ouimet advised him the Employer intended to commission Willis & Associates to do the work on behalf of the Treasury Board. On December 19, 1989, Willis wrote to Ouimet declining to conduct a further analysis "unilaterally" on behalf of the Treasury Board. Willis testified he understood from the very beginning he was answerable only to the JUMI Committee. Willis felt this was inappropriate. Willis had hoped the JUMI Committee would reconvene. He was asked by a Treasury Board representative, Gaston Poiré, under what circumstances he would conduct the analysis. Willis responded that he would conduct a study of a larger sample if the Commission requested it, since "the Human Rights Commission was an objective third party and it was their bill." (Volume 59, p. 7311).

522. In Willis' letter to Ouimet of December 18, 1989, he mentions for the first time what the information from a second study should provide. The relevant portion of the letter reads:

> It is my belief that an expansion of this analysis is necessary to determine the extent of any actual bias that may exist in the evaluations. This information should afford a basis for any adjustment in evaluation results that may be required to assure a fair and objective study. [emphasis added]
> (Exhibit HR-92)

523. On January 23, 1990, the Alliance announced its permanent withdrawal from the initiative and three days later, on January 26, 1990, the President of the Treasury Board announced the implementation of equal pay for work of equal value adjustments, with the assurance the

government's action did not prejudice any conclusions and findings of the Commission relating to the resolution of the issues still to be investigated by the Commission.

524. Brookfield testified she noticed a change in the attitude of the Employer toward the end of the study. She made reference to the fact the discussions between the unions and management was initially about apparent gender bias. Following the Wisner 222 report however, the Treasury Board no longer discussed apparent gender bias and had changed their approach by suggesting they would adjust for actual gender bias.

525. The unions were very concerned about this change in the Treasury Board's approach after the Wisner 222. Brookfield testified there was correspondence about adjusting scores and referred to a letter written January 26, 1990, after the break down of the study (Exhibit HR-41), from the President of Treasury Board to Max Yalden, Chief Commissioner of the Commission, explaining the equalization payments were calculated on the basis of adjustments for gender bias made by Treasury Board.

526. In the letter, the President of the Treasury Board, Robert de Cotret, wrote to Yalden with details of the government's decision to implement service wide measures based on the evaluation results of the Joint Initiative. The letter does not refer to the extent of apparent gender bias identified in the Wisner 222, but instead alludes to "the extent of gender bias." An excerpt from de Cotret's letter reads as follows:

> It is my strong belief that an unprecedented study of this magnitude must be fair, statistically sound, and credible, given its significant ramifications. This further analysis was needed to determine the extent of gender bias and adjust the Initiative's evaluation results accordingly. I appreciate, therefore, the Commission's agreement to conduct this analysis to determine the extent of gender bias. [emphasis added]
>
> (Exhibit HR-41)

527. The above excerpt seems to confirm the union's belief of the changing emphasis by the management side from a concern for apparent gender bias raised in the Wisner 222 to an issue of adjusting results to account for actual gender bias. Brookfield testified it appeared to her the Treasury Board had made a decision there was definitive evidence of gender

bias in the Wisner 222 and all that needed to be done was to adjust the scores for the bias.

528. In early 1990, Willis was contacted by the Commission. This contact was made after the Alliance had announced their withdrawal from the JUMI Study. Willis was informed by Durber that the Commission had determined an additional analysis was necessary based on re-evaluations to be undertaken by Willis & Associates. The Commission itself would, however, analyze the results of the Willis re-evaluations.

529. In Willis' opinion, the only alternative to a further study, would be to use some other evaluation system which would have, in effect, reconstructed much of the study. This exercise would have been extremely costly. Willis also expressed his opinion as to what ought to be done with the study results. He suggested the Tribunal has three alternatives: (i) to implement the study as it is; (ii) to adjust the results; or (iii) to trash the study. Willis maintained he would rule out trashing the study, and would adjust the results for any possible gender bias.

E. THE COMMISSION

530. When the Commission responded, in April of 1985, to the invitation of the President of the Treasury Board to support the JUMI, the Commission agreed to put on hold the investigation of s. 11 complaints filed prior to the announcement of the JUMI, as well as complaints filed subsequently to the announcement of the JUMI. The Commission indicated it would await the results of the study before taking action. This also depended upon the circumstances at the time of the filing of the complaints.

531. The Commission's response to the invitation was contained in a letter dated April 17, 1985 (Exhibit HR-18, Tab 18), from Gordon Fairweather to the Honourable Mr. de Cotret. That letter indicates that if the Commission satisfied itself the methodology employed in carrying out the study was consistent with s. 11 of the Act, then it would issue a special guideline advising that the study was consistent with the Act. It would also issue guidelines for the implementation of corrective action in accord with s. 11.

532. The Commission participated in the JUMI Process only as an observer. Representatives of the Commission attended the JUMI Committee meetings and when asked by members of the JUMI Committee provided clarification and advice relative to the JUMI Study. Participation by the Commission was mainly of a technical nature, and involved such tasks as selecting samples in the ICR testing and dealing with problems of interpretation relevant to the Act and Guidelines. Commission employees

also attended as observers during the operation of the five and nine evaluation committees.

533. The Commission did not intend to be a party to settlements reached by the parties to the JUMI. It did, however, intend to examine any agreement reached to determine whether it met the requirements of s. 11 of the Act.

534. In early May, 1989, Durber joined the Commission as Chief of Equal Pay. This title was later changed to Director of Pay Equity. On June 12, 1989, Durber met the JUMI Committee co-chairs and expressed his concern that if the parties were unable to determine what should be done with the Wisner 222, the initiative could easily founder. Durber testified the co-chairs agreed at this meeting that all the parties, including the Commission, ought to have free access to the job evaluation results from the JUMI Study.

535. Durber advised the co-chairs at that time the question for the Commission was how to interpret the job evaluations that had been done. He emphasized if there was gender bias the Commission would have to be involved because it needed to know whether the evaluations were acceptable as evidence, should the Commission pursue the complaints filed by the Alliance.

536. No formal investigation of the complaints was done by the Commission until March 6, 1990. On that date, at the request of the Commission, the JUMI participants met with the Commission to review outstanding issues. By that time the JUMI had permanently broken down.

537. The next significant date is March 6, 1990, when the Commission met with the JUMI participants. The Commission wanted to reduce the number of issues arising from the JUMI should the complaints be referred to a Tribunal. The Commission's press release, following the meeting, specified the Commission must be satisfied that all the requirements of the Act had been met. It also specified Treasury Board had given the Commission the calculations used to predict their adjustments which the Commission would examine in its investigation.

(i) Commission Investigation

538. When the JUMI Study ended in the beginning of 1990, it became evident to the Commission its role as observer in the JUMI Study was also at an end and it was time to begin pursuing the normal complaint process.

The question of apparent gender bias raised by the Wisner 222 was a part of the investigation into the complaints. The approach by the Commission was to treat the question of apparent gender bias as the first focus of its investigation into whether wage discrimination persisted in the Federal Public Service. The government had made equalization payments in January, 1990, and the Alliance maintained those payments had not closed the wage gap, leaving wage discrimination still in place.

539. Gender bias was a consideration when the President of the Treasury Board announced the wage equalization payments in January of 1990. The Treasury Board President had not indicated the extent to which the equalization payments accounted for the bias, but did state in his announcement the Commission would be examining the matter.

540. The Commission's approach to the investigation as described in Exhibit HR-55, "Notes for Presentation on Alleged Gender Bias in Job Evaluation of the Joint Initiative" was conservative in terms of the amount of evidence it sought for in addressing the question of apparent gender bias.

541. Durber testified the Commission investigated all five complaints from both the Alliance and the Institute simultaneously. It was probably the speediest Commission investigation performed prior to that time because the Commission had before it all the job evaluation data gathered from the JUMI Study. The Commission had no need, therefore, to conduct its own job evaluations.

542. There were four areas for investigation by the Commission. The first involved the investigation of gender bias. The Commission had to decide whether they could rely on the job assessment information from the JUMI Study. The second involved looking at any wage gaps that might appear. The Commission had to develop a methodology to calculate wage gaps. The third area for investigation involved considering and valuing benefits. Finally, the fourth area, (not yet complete), involved parts of two complaints which bore on limitations on employment opportunities as a result of compensation practices.

543. An overview of the chronology begins with the Commission's investigation starting in March, 1990, arriving at tentative conclusions on gender bias in July of that year. In the same month, the Commission briefed the parties on its findings regarding "apparent gender bias" in the committee evaluations. In August, 1990, the Commission produced a draft

report on the wage gap and the parties were briefed on the Commission's interim findings regarding its conclusions.

544. There was also a meeting in August with the parties on the status of the Commission's investigation pertaining to the valuation of benefits. In September, 1990, the Treasury Board submitted a written response to the Commission's August draft report. The final investigation report went to the Commissioners in late September, 1990. The following October, the Commission made its decision with respect to the wage gap on the five complaints and requested the President of the Human Rights Tribunal to appoint a tribunal.

545. The Commission's investigation into the s. 11 complaints is contained in Exhibit HR-250, entitled, Investigator's Report: Wage Adjustment in the Federal Public Service - Possible Gender Bias in Job Evaluation Data. Durber released the Investigator's Report on this subject to the parties in September, 1990. The report contains the Commission's findings and conclusions relating to the question of apparent gender bias in the committee evaluations. The Commission's conclusions are found in para. 51 of that report which states as follows:

51. Conclusions

> Commission staff have found that the Willis checks reveal some differences between consultants' evaluations and those performed by the Joint Initiative. Investigators do not find that these differences reveal patterns that can be correlated consistently with gender or occupation in the Joint Initiative evaluations. The extent of possible "undervaluation" of male jobs is less than 3%, but can likely be accounted for by differing understandings of work described, as well as the meaning of bench marks and the application of the Willis plan. It is not apparently the result of bias linked to sex. Moreover, the 3% is not evenly distributed across occupations. Certainly, the fact that two sets of independent groups (Willis consultants and the Quality Analysis Committee) could produce results varying by a margin of 2% to 3% indicates that such differences may be expected and be due to reasons other than bias.
>
> (Exhibit HR-250, Part I)

546. Durber's oral evidence corroborates and confirms the contents of this report and focuses on the steps pursued by the Commission in

investigating the possibility of gender bias in the committee evaluations. It is noted from Durber's evidence, further testing procedures were undertaken by the Commission subsequent to the commencement of this hearing. Both the investigative procedures conducted as part of the Commission's initial investigation and the subsequent testings conducted at the request of the Commission will be reviewed by the Tribunal.

547. The Investigator's Report indicates there was no clear evidence of gender bias in the evaluation results. The report contains a recommendation of formulae for equalizing pay between males and females which pay ought not to be adjusted for possible gender bias. It proposed that the Commission accept its findings vis a vis the related complaints under s. 11.

548. A draft of the Investigator's Report (Exhibit HR-250), was provided to the parties for comment in the summer of 1990. The Treasury Board responded by letter and written report dated August 17, 1990, from Ouimet in her capacity as Assistant Secretary, Classification, Human Resources Information and Pay Division, addressed to Durber. Ouimet testified during the voir dire hearing of the Tribunal but was not called when the hearing reconvened. The last paragraph of her letter concludes that the Commission's investigation was deficient and did not demonstrate a clear case there was no gender bias. On the other hand, she expresses the view as to the unlikelihood of any party being able to demonstrate the existence of gender bias in the results. The paragraph is reproduced as follows:

> On the other hand, it is unlikely that anyone could demonstrate gender bias does exist given that the Willis firm has not provided a baseline by which evaluation results may be compared from study to study. It is not possible to measure adequately the application

> of the plan so as to conclude definitively that bias does or does not exist. Do not conclude however, that we should not examine very closely all the rating inconsistencies raised by the various committees of the Joint Initiative, your own research, and ours. It is now vital that we leave aside the `why' behind rating anomalies and focus instead on how they may be corrected. We would be prepared to contribute to the design of an appropriate study to resolve rating inconsistencies. [emphasis added]

(Exhibit HR-46, p. 2)

549. In the detailed comments attached to her letter, Ouimet asks the rhetorical question, "Is it possible to distinguish between evaluation biases along sex lines and the overall application of the Willis Plan in a manner that would assign an appropriate weight to each?" The report states the answer must indicate the degree to which the question of gender bias is purely a statistical or substantive question. In the latter case, according to Ouimet, statistics may contribute little.

550. In the Treasury Board's written response to the Commission's final report on Possible Gender Bias in the Evaluation Data, which is contained in a letter from Ouimet to Durber dated September 7, 1990, the Treasury Board is clearly of the opinion a statistical study is not the best approach when determining possible gender bias. The following excerpts from her comments at p. 1 of the report are helpful in understanding the Employer's response:

> In essence our disagreement can be summarized as follows: the Investigator embarked on a highly restricted look at gender bias through statistical research that was inappropriately conducted. Even if it were appropriate, the restricted nature of the overall study is such that nothing can be said about the issue of gender bias since the important issues implied by it were never examined.
>
> The Commission quotes at length the position of the Public Service Alliance of Canada (PSAC) that many of the issues are non-statistical. We are in agreement with this position and have argued that statistical analysis in this area is useful only insofar as it may raise the possibility of a problem that would require a non-statistical approach to answer. Notwithstanding this objection, we are of the opinion that any statistical study, no matter how adequate, is not the best approach in this matter. There is so much judgement involved in the scoring of any job questionnaire that to determine gender bias statistically is difficult at best because it requires that a weight be assigned to every factor of judgement/bias/inconsistency, what have you, to the score itself. Since you have decided to restrict your study to a statistical analysis of Willis evaluation data, we feel compelled nevertheless to critique your study on statistical grounds.
>
> The long critique we sent to you was an attempt to demonstrate, through statistical arguments, that the approach taken and the

> empirical findings do not, under any circumstance, permit you to conclude with certainty there is no gender bias in the Joint Initiative evaluations. The most you can conclude is that there is not enough evidence to decide one way or the other. You have not addressed any of our concerns systematically other than through an editorial comment that our 'statistical criticisms make rather too fine a point'.
> (Exhibit HR-250, Tab J, pp. 1-2)

551. The Treasury Board apparently had used an alternative line of enquiry into the question of possible gender bias described in Ouimet's detailed comments of September 7, 1990. Using a different approach, she writes the Treasury Board came to the same conclusion as Sunter, but in their view, the conclusion is misleading since it only represents half the story and says nothing about how often questionnaires are under- or over-evaluated. The Treasury Board's overall conclusion is found on page 10 which states:

> Using criteria provided by the Willis firm, it is not possible to conclude that while there may be statistically significant differences in patterns of evaluations, they are not substantively important. As shown above, the issue of level of difference has ignored the frequency dimension and the differences in patterns are indeed significant. We attempted to take into account mis-evaluations in order to see whether there was a gender pattern to them and it would appear there is.
>
> We have analyzed the same data and using the same measure as the Sunter analysis, and yet reached different conclusions. We are convinced that the data show serious problems with the evaluations and that these problems look very much like gender bias; in any event, further analysis is required. We remain firm in our belief that the scores need to be adjusted, but we are prepared to discuss a different adjustment strategy from the one originally used. Any adjustment is going to be difficult to estimate given the significant differences between the two Willis studies. [emphasis added].
> (Exhibit HR-250, Tab J)

552. We will now describe and examine specific factual information found in the Commission's investigation provided by Durber. On March 8, 1990, the Commission received from the Treasury Board, a document (Exhibit HR-185) which explained the methodology used by the Employer in making its equalization payments. According to Durber, the Treasury Board paper, issued in March, 1990, estimated an average bias of +3 per cent for evaluations of positions from female-dominated occupations and of -4 per

cent for evaluations of positions from male-dominated occupations. Accordingly, the wage equalization payments had therefore incorporated a corresponding across-the-board adjustment when calculating equal pay for work of equal value. The adjustments resulted in payments to public service employees in female-dominated occupational groups which were lower than they would have been without those adjustments for possible gender bias.

553. The revision of scores is explained in the methodology paper as follows:

> A score revision factor based on simple statistical techniques was estimated by the Treasury Board. All questionnaires except those rated by the Master Evaluation Committee and the Willis consultant were revised: ratings for female questionnaires were reduced by approximately 3% overall and male questionnaires were raised by roughly 4% overall. All policy analyses presented in the remainder of this report use the revised evaluation scores as described.
>
> (Exhibit HR-185, pp. 6-7)

554. In attempting to understand Exhibit HR-185, which contains a good deal of detailed statistical jargon and information, Durber sent the report to seven independent individuals for their comments. These individuals included pay equity experts, Weiner, Dr. Morley Gunderson, Lois Haignere, Willis & Associates, Roberta Rob, Judith Davidson-Palmer, and a statistician, Sunter. Durber viewed these individuals as potential participants in a workshop the Commission had scheduled for April, 1990, to review the Treasury Board's methodology (Exhibit HR-185) and to advise him how he ought to deal with it.

555. The Commission had difficulty in obtaining data from the Treasury Board during its investigation of the complaints. Durber testified the actual data the Treasury Board used to arrive at its conclusions in HR-185 were never produced. The Commission had to project salaries and create their own salary data bases because of the length of time it took the Treasury Board to provide salary information. A complete set of the salary data was finally provided to the Commission during these hearings.

556. On April 9, 1990, the Commission held its workshop and some of the individuals who are listed above attended, namely, Sunter, Roberta Rob, Judith Davidson-Palmer, and a representative of Willis & Associates. The

others, who did not attend the meeting, provided written comments. Durber wanted to be "as well informed as possible by some of the better minds in Canada on the issue of pay equity." (Volume 147, p. 18197). After fairly extensive consultation with these individuals, Durber consolidated the advice he received and formulated an investigation plan and hypothesis.

557. Following the meeting of April 9, 1990, Durber consolidated the advice resulting from his discussions with these individuals in order to clarify the issues needed to be addressed by the Commission. A decision was made to challenge the Treasury Board methodology by detailed questioning.

558. The Commission was also interested in knowing whether the factors in the Willis Plan, were different from the results for male-dominated occupational groups as opposed to the results from female-dominated occupational groups. Durber contracted the Wyatt Company, an international company of management consultants which enjoys a considerable job evaluation practice. The Wyatt firm was asked to use the database for all

the JUMI Study job evaluations. The Wyatt firm looked at the data to determine whether the relationship between the factors was the same regardless of the gender of the group and regardless of the occupation from which the questionnaire was taken. Their report was provided to the Commission in early June, 1990. The Wyatt analysis demonstrates there were correlations between various factors, for example, the extent to which a score on mental demands correlates with knowledge. The conclusions from this report was there appeared to be no significant differences in the correlations between the factors for the male and female jobs or between the overall patterns. The report further indicated there was some difference in scores on working conditions between male and female jobs. It was Durber's belief this was explainable by the nature of the work. (Volume 147, p. 18208).

559. The approach of the Commission in assessing gender bias was not "to prove no bias" but simply to find whether or not a reasonable person would see bias operating. (Volume 149, p. 18521). According to Durber, because there is a different pattern for males as opposed to females, as for example in the Wisner 222, this does not tell the investigator anything except, perhaps, "whether one ought to look further".

560. During the initial investigation, a letter dated June 20, 1990 accompanied by a binder, was delivered by a representative of the Treasury Board to the Commission, which contained information relevant to the

Commission's assessment of gender bias. The documents in the binder included IRR Sub-Committee documents, the ICR studies, the recommendations for changes prepared by the Willis consultant, Drury on the MEC evaluations, the Tristat Report, Willis' report on MEC's work dated July, 1988, questions referred to Willis in August, 1988 from the management side regarding the MEC evaluations, minutes of JUMI Committee meetings, copies of letters written in July, 1988 by the Alliance and the Institute to Drury regarding evaluation rationales and interpretation of the factors under the Willis Plan, Willis' response to committee challenges of the MEC evaluations, a copy of a letter from Willis regarding Committee #4 written on August 17, 1989, and copies of letters between the parties regarding the Wisner 222.

561. Durber stated the documentation in the binder provided by the Employer did not particularly pertain to gender bias. In the set of documents relating to the ICR Sub-Committee, Durber searched for specific evidence of gender bias. With regard to the Tristat Report he testified he was looking for bottom line conclusions because he wanted to know whether in fact there had been indications, or hard evidence of gender bias. As to the ICR studies, considering the small number, 25 tests, it was not possible, he said, to detect a trend.

562. Durber spoke with one of the Commission's observers, Brian Hargadon, concerning his observations of the ICR tests. Hargadon participated in all of the tests. He also administered some of the tests. Durber testified that in Hargadon's view the evaluation committees did not, over time, take the tests as seriously as when they had begun. Durber, therefore, found the ICR tests inconclusive on the question of gender bias.

563. On the question of the changes to the MEC evaluations prepared by the consultant, Drury, Durber primarily relied on Willis' opinion that the matters brought forward by her were resolved. As to the report prepared by Willis & Associates in July, 1988 and their analysis and conclusions regarding the MEC's work, Durber considered the bottom line in the report to be that there was no problem with gender bias. In Durber's opinion, after reviewing the materials submitted to him by Treasury Board, he came away with no better understanding of how gender bias might operate in the job evaluation results. Durber testified the Treasury Board material was not helpful and he needed to better understand whatever was going on with respect to so called gender bias. Accordingly, he decided to look elsewhere for answers.

564. Durber stated the only discussions he had with Treasury Board staff about material contained in the binder was during a presentation he made to the Employer on July 5, 1990, regarding issues surrounding gender bias. A more detailed analysis of gender bias as viewed by the Treasury Board was not made available to the Commission until August, 1990. It was then the Treasury Board submitted its more detailed written submission concerning this subject to the Commission.

565. Part of the Commission's investigation into the question of apparent gender bias was a follow through of the recommendation, contained in the Saveland Report, for further analysis of the 27 "under-evaluated" male jobs (subsequently reduced to 25). These jobs had been identified by Saveland as showing a difference of 10 per cent or more between Wisner and the evaluation committees. Durber convened a joint committee in the spring of 1990, composed of management and union employees under the chairmanship of Ron Renaud, Senior Consultant, Equal Pay Section of the Commission. They met for two weeks beginning on April 30, 1990. In the Commission's letter to the committee members, the committee was informed as follows:

> The committee's mandate is to carry out a quality check of twenty seven positions that were evaluated by JUMI committees coming after MEC. In an analysis of 222 position evaluations by Willis and Associates, June, 1989, it was found that the evaluations were significantly different from the MEC discipline and contributed most to the finding of apparent gender bias.
> (Exhibit PIPSC-135)

566. Former MEC evaluators were selected to participate in this committee, including two management and three union representatives whose names were suggested by the Employer and the unions. Durber wanted participants who had a breadth of views. In his opinion, this goal was achieved. This committee was referred to as the Quality Analysis Committee (the "QA Committee") and produced a report, The Quality Analysis Report.

567. Durber testified that within the context of the QA Committee, he was less interested in the fact there were differences between the consultant and the multiple evaluation committees, than he was on what accounted for these differences. He was interested in knowing whether the QA Committee members perceived the multiple evaluation committees and consultant differences in a way that related to the fact these were male

jobs, or whether they perceived any bias on the part of the multiple evaluation committees. He considered the five former MEC members to have a

special insight into both the Willis Plan and the MEC discipline. He expected they would understand "the mechanisms behind their own differences with the committees."

568. Durber testified the Commission was trying to determine if there was a reason, a motive, or some conscious or unconscious effort by the multiple evaluation committees to disfavour these male jobs. If gender bias was to be evident anywhere, he reasoned, it would be evident with these 25 jobs.

569. The procedure followed by the QA Committee in completing its assignment was for each committee member to read the questionnaire, independently evaluate the questionnaire, review the MEC benchmarks used by the JUMI Committee and those used by the Willis consultant, and then select additional appropriate MEC benchmarks.

570. The evidence before the Tribunal is contradictory as to whether or not the QA Committee was required to arrive at a consensus in their evaluations. According to Durber's evidence, the QA Committee were not asked to form a consensus. Durber testified the Commission asked each QA Committee member to report to the chair on their evaluations, then discuss them, but not arrive at a consensus. Durber further testified the Commission was not attempting to validate the ratings of the 25 jobs but simply wished to understand whether the members of the QA Committee might become aware, during this process, of any gender issues either in their own ratings or in the multiple evaluation committees' ratings.

571. On the other hand, two union members of the QA Committee testified the QA Committee was asked to reach a consensus and failed to do so. Their evidence is the QA Committee followed the same Willis procedure used by the evaluation committees. The only exception, according to these witnesses, was that the consensus had to be unanimous for each sub factor in the Willis Plan, rather than the two-thirds majority required for consensus in the evaluation committees. An attachment to the letter dated April 23, 1990, from the chair to the QA Committee members corroborates and confirms the unanimous agreement requirement for consensus. The relevant part of that document states:

> Evaluation findings will be arrived at by committee consensus. This means that the evaluations by factor, sub-factor and points must be agreed to by each member of the committee.
> (Exhibit PIPSC-135, p. 3)

572. Durber testified that at the conclusion of the QA Committee's work, the chair of the Committee, Ron Renaud, reported to him the differences in the ratings between the QA Committee and the evaluation

committees were due to "perceptions of the work", but that the QA Committee found the gender of the jobs played no role whatsoever in the ultimate evaluations. A review of the written report does not include any reference to this verbal report from Renaud to Durber.

573. Durber concluded from this exercise, it would not be unusual to find a range of views between evaluators which would be reflected in a range of ratings. Durber interpreted the difference between Wisner and the evaluation committees as "normal, honest disagreement about work as opposed to any problems with gender bias."

574. Durber stated in his evidence, "that the entire edifice of the question of gender bias which is before the Tribunal rests on a foundation of one person's view [i.e., Wisner's] of 25 questionnaires." (Volume 149, p. 18581).

575. Durber used the QA Committee Report to compare the average of the QA Committee evaluator's ratings to the total point score given by the JUMI evaluation committees and Wisner. According to Durber, this comparison indicated to him the QA Committee disagreed as often with Wisner as with the evaluation committees and he states in Volume 149, at p. 18573, line 14 to p. 18574, line 22:

> The patterns were that the low raters agreed, essentially, as often as they disagreed with the committee ratings.
>
> The high end rater agreed only about one-third of the time with the committees, although a third of the time was still a reasonable number.
>
> I concluded from this exercise that in fact one should expect a range of view, a range of ratings on jobs, that it wasn't unusual to find a range of ratings, that it certainly would not be unusual to find differences between any raters.
>
> That permitted me to believe, interpret Mr. Wisner's differences from the committees as a normal, honest disagreement about work as opposed to any problems with gender bias.
>
> The fact that they were male jobs may or may not have been coincidental, but I could not see any necessary reason to believe that there was bias operating as a result of the differences between Mr. Wisner and the committees.

I didn't, for example, conclude that Mr. Wisner was biased in favour of male jobs, which could have been one of the interpretations from his report. He being a male, one might have concluded that. But whether he was a professional consultant and objective or whatever was another issue.

But we did find these five individuals from MEC also disagreed less than Mr. Wisner, but probably about as often or a little more than Mr. Willis when he and his other three consultants had looked at male jobs.

576. One of the union representatives on the QA Committee, Tim Yates, was asked in chief about his understanding of the purpose of the QA Committee. His response was that its purpose was to look at the committee

evaluations, ascertain if they had chosen appropriate benchmarks and correctly applied them. Yates testified he could not recall any instances, during this review, where inappropriate benchmarks were used. As to differences between the consultant's evaluations and the committees' evaluations, Yates says the following in Volume 175, at p. 22226, lines 4 - 22:

A. Well, if one is to make a huge assumption, that we were the experts in the thing, sometimes we were higher than the consultant, sometimes we were lower than the committee. I think it was Mr. Willis who said many times, "this is not a science".

I would say personally that what was the problem? It all appears to be within tolerance.

Q. What do you mean by it would all appear to be within tolerance? Where did that phrase come from?

A. The lowest possible difference is one step. One step is 15 per cent. That's the very slightest possible bit of shading in any factor is 15 per cent.

577. The other union representative who testified regarding the QA Committee was Mary Crich, who had been an alternate on the MEC and participated in the committee evaluations as a member of Committee #5. She was asked about her observations. On reflection, she found her participation on the QA Committee was a good experience because it led her to understand that what she had done as a committee member was precisely

what the evaluation committees were supposed to have been doing. She found the QA Committee evaluations were reached by exactly the same discussions relating to the same points and with more or less the same kinds of agreements and disagreements she experienced in her evaluation committee.

578. As to Crich's understanding of the work of the QA Committee, she testified the individuals selected for the QA Committee knew the MEC discipline, and thus could decide whether or not the ratings of the evaluation committees respected the MEC discipline or differed significantly. Crich further testified when the QA Committee finished its work, there was general agreement among the members there was no bias. If there was a significant difference, it was, according to Crich, because it was a genuinely difficult job to evaluate which had no comparable benchmark. She described the 25 jobs as "very difficult jobs". Crich was asked in cross examination what she understood was meant by "bias". She responded in Volume 192, at p. 24830, lines 5 - 15:

> A. What I remember the other participants saying is that there had been allegations in the media that there had been -- the results of the study were biased and by "biased", that meant that the evaluations had not been fair to all jobs equally and that female jobs had been rated too high. I don't know if the -- also that male jobs had been rated too low. Maybe it was both. Maybe it was just one or maybe it was -- but that was -- the bias is that female jobs were rated too high.

579. A further clarification of this response was given by her in Volume 192, at p. 24841, lines 2 - 12:

> Q. Mr. [sic] Crich, I just have one question, really, and I will try to phrase it as clearly as I can.
>
> When your Quality Assurance Committee agreed that there was no bias in these questionnaires, these 27 questionnaires that you evaluated, were you looking at the reasons for the difference and, therefore, concluding that the gender of the questionnaire was not the reason for the differences?
>
> A. That's correct.

580. Willis testified he had a number of problems with the QA Committee. He was disappointed with the composition of the committee and would have preferred if the total MEC membership had been reconvened rather

than only the five individuals selected. Another factor which troubled him was although three of the members were from the MEC, two of them had acted only as alternates. Moreover, one of the members had been identified as an outlier in the Tristat Report. Willis also believed, since two of the members had participated in the evaluation committees, their opinion about the committee results might be suspect.

581. Another area of concern for Willis was these five individuals had not done any prior evaluations for at least two years. Willis testified this committee should, at the very least, been given a day or two of refresher training by the consultants. In his opinion, it would be difficult after a two year lapse in time to return and do evaluations, particularly evaluations which were to be critiqued. His biggest concern is noted in Volume 208, at p. 26950, lines 3 - 8:

> However, I guess my biggest concern about the QA committee was it was my understanding that there was no consensus process. To me I look at the consensus phase of the evaluation process as being part of the data-gathering collection.

582. Willis stressed the consensus phase of the Willis Process is a very important exercise because it gives the committee members opportunity to discuss the facts of the job and time for all members to consider the information thus elicited. It is the "fine honing of the information" which is important, according to Willis, in this stage of the process when committee members change their evaluation at this point, Willis believes the change is appropriate as long as it is based on facts which are brought out as a result of the discussions. On this basis, Willis discounted the results of the QA Committee because an essential and critically important step was left out. Willis testified that to some extent he might change his opinion regarding consensus, if indeed the QA Committee included the consensus process in their deliberations.

583. Durber testified, in the normal course of an investigation, the Commission expects an employer to provide evidence in support of "their defence". He testified the Commission receives a defence from the employer

which says, in effect, it ought to be excused from accepting the results of its own study. Notwithstanding, the Employer was duly represented, and presented no evidence on which to support such a conclusion.

584. According to Durber, differences between a committee and a consultant are bound to occur, but the Commission needs to be vigilant about understanding those differences and their relationship to gender.

585. The Commission opted to conduct further analysis of the consistency of the evaluations by the nine evaluation committees compared to those of the MEC. Durber felt he had no alternative but to order another study so as to complete the picture. He was not happy with the alternative because, in his opinion, it was impossible to replicate job evaluations done by the committees. Durber felt uneasy about the validity of the process, which he described as people in a sense second guessing what a rather large number of people had done over a period of time. Durber would have preferred to have the parties to the JUMI Study deal with the issue of apparent gender bias in their own way. He elaborates in Volume 149, at page 18599, lines 1 - 11:

> But conceivably they might well have had committees explain their results, look at the differences between themselves and Mr. Wisner. There might well have been some judgments raised or brought to bear on the patterns themselves and on the differences between the committees and Mr. Wisner.
>
> There could have been some good rationalization, if you like. But in the event, that proved not possible. Once the committees were gone, they were gone.

586. Durber said in the course of his investigation, he did not contact Wisner because he preferred to relate to what he considered "reasonable criteria for judging the quality of job evaluation." The issue, in Durber's view was one of differences between committees and consultants and the process followed by the committees. Durber questioned why he should prefer to believe a consultant over the evaluation committees. Given a choice between the judgment of a group of people who are well informed as opposed to following the discipline of one individual, Durber would prefer to believe the group of people. This was one of the "indirect measures" which Durber used in drawing his conclusions about gender bias. Durber believed if he contacted Wisner, he then would have been bound to call each of the working committee members.

587. Durber contacted Willis to do a further evaluation in the early part of 1990. Willis confirmed his acceptance by letter of February 12, 1990 to Durber which states in part:

> The purpose will be to determine the extent of any systematic bias that may exist in the results of evaluation committee efforts.

...the sample size should be 300 positions, with at least 131 being from male dominated occupational groups and the balance from female dominated occupational groups.

127

As to the sample selection, the random selection methodology we used in the earlier special analysis would, I believe, be acceptable to the unions and management. The Human Rights Commission should have input into this methodology...

The method employed for the analysis will be the same as used in our previous analysis. Each selected questionnaire will be reviewed and a determination made as to whether a similar position is included among the Master Evaluation Committee's evaluations. In cases where a similar MEC benchmark exits [sic], the MEC evaluation will be adopted as the consultants evaluation. When no similar benchmark exists, the consultants will do an independent evaluation of the position, supported by reference to appropriate MEC benchmarks. Comparison will then be made with the sample committee evaluation and rationale for that position. When differences are found between the consultants evaluation and that of the committee, a written rationale explaining the consultants evaluation will be provided.
(Exhibit HR-93)

588. Durber stated his objective in commissioning Willis to re-evaluate the additional 300 positions was essentially to pursue the issue which had been raised as a result of the Wisner 222 relating to possible gender bias. In view of s. 9 of the Guidelines, supra, Durber wanted to be assured there was no question of gender bias. He further testified he could see no alternative but to pursue the same approach as Wisner had because it was through that approach the issue had arisen in the first place.

589. Durber would have preferred to engage Wisner to perform the second set of re-evaluations but, in the meantime, Wisner had left the Willis firm. Accordingly, Willis was authorized to form a committee consisting of four consultants, (collectively referred to as the "Gang of Four"), who were to perform the 300 re-evaluations (the "Willis 300").

590. Willis testified he understood there was a concern the four consultants working together would arrive at a slightly different result than Wisner. Accordingly, their additional task involved selecting jobs from among the Wisner 222 and independently evaluating them without making

any judgment as to differences between the Gang of Four's and Wisner's re-evaluations. Using the Gang of Four, Willis was to review approximately 20 per cent of the evaluations of the Wisner 222, i.e., 44 questionnaires, as a double check on Wisner's interpretation of the jobs.

591. The Gang of Four tried to match, as closely as possible, the methodology that had been used in the Wisner 222. The sample of positions was selected by the Commission and was taken from the total sample of evaluations excluding the MEC evaluations and any re-evaluations included in the Wisner 222. Willis was not asked to do any analysis of those re-evaluations. Once the Gang of Four completed the 300 re-evaluations, the results were turned over to the Commission for analysis.

592. The Gang of Four consisted of Willis, two of his associates, Owen and Davis, and one outside bilingual consultant, Esther Brunet. Questionnaires were assigned to each consultant and a second consultant reviewed each of those evaluations, so that there were always two consultants involved. The work took approximately two months. A report entitled Report to the CHRC Equal Pay, Quality Analysis of Sampled Committee Evaluations, Joint Initiative Equal Pay Study, was presented by Willis to the Commission in March of 1990.

593. Although this review was to assess the quality of the Wisner 222, Willis & Associates were instructed by the Commission not to draw conclusions as to the quality of either their work or Wisner's re-evaluations. In the course of these hearings, and in the context of this review, Willis was asked his opinion on the quality of the Wisner re-evaluations. He replied in Volume 59, at p. 7337, lines 11 - 24:

> THE WITNESS: I was satisfied with the quality of the Wisner evaluations six or seven months earlier when I looked at his rationales and I looked at his actual evaluations. I had a great deal of confidence in Mr. Wisner's ability as a professional job evaluator.
>
> I did not, at this point, sum up the 44 evaluations by our team of consultants and compare them in total with the Wisner evaluations. They were not identical, there were some differences. But I felt that it was up to Mr. Durber to analyze those differences and, in effect, decide whether the quality was consistent between both consultant teams.

594. In terms of analyzing the results of the 300 evaluations, Willis stated it would have been appropriate, in his opinion, to perform a statistical analysis to identify the existence or non-existence of a systematic pattern of gender bias. Had the Commission asked Willis to perform this analysis, he would have retained a statistician, Dr. Milczarek, who in the normal course of events, performs this kind of analysis for him.

595. The last communication between Willis & Associates and the Commission, concerned the 44 re-evaluations. This took the form of a letter dated May 1, 1990, written by the Willis consultant, Keith Davis to the Commission. During the re-evaluation of the 300 positions and the review of the 44 Wisner re-evaluations, the Gang of Four inadvertently referred to a list relating to the working conditions factor in the Willis Plan. Changes had been made by the JUMI Committee to this factor which the Gang of Four had failed to take into account. Davis informed the Commission when using the re-evaluations, the working conditions factor needed to be changed. In the end, one re-evaluation by the consultant required a change.

596. The Tribunal had the benefit of hearing evidence from Esther Brunet concerning her participation in the re-evaluations as a member of the Gang of Four. Brunet was the only member of the Gang of Four who was an employee of the Federal Public Service. She had been involved in the

JUMI Study as a chair in the first version of Committee #4. Her employment background at the relevant time was Director of Personnel, Finance and Administration with the Status of Women Canada. Willis testified he needed a French speaking consultant to participate in the Willis 300 and, because he and his staff had a great deal of confidence in Brunet's ability to evaluate, they contracted with her to evaluate the French questionnaires.

597. Brunet evaluated approximately 100 questionnaires out of the total of 300. About 70 per cent of those were French questionnaires. She first evaluated the questionnaires independently. If the evaluation committee had used only one benchmark, she would try to find more. Once her evaluation was done, she would look at the evaluation committee scores and rationales, and if she felt the reason for the difference made sense, she would give the benefit of the doubt to the evaluation committee scores, if not, she would then prepare her justification and present it to the other three consultants. During this presentation, Brunet would try to convince the other three team members of the need for the change she was proposing. If she was unable to persuade the other members, the evaluation

committee scores remained as they were. Brunet explained the Gang of Four did not write rationales in the same manner as the committees because the reason they wrote them was simply to justify the difference between the consultant and the committee.

598. Brunet's evaluations of the French questionnaires can be compared to the committee scores because she was the only consultant in the Gang of Four evaluating French questionnaires. The French questionnaires are summarized in PIPSC-162 and confirms that for female-dominated questionnaires, Brunet's average score was 157.1 compared to the committees' average score of 157.9. With respect to the male-dominated questionnaires, her average score was 250.7 compared to the committee's average score of 249.7. Brunet rated the same as the committees except in eight cases, five from the female and three from the male.

599. Unlike other Commission investigations, the investigations here under s. 11 differed somewhat, in that the factual foundation for the complaints were known to the Commission because it had participated in the process as observer from an early stage. The Commission observers attended the JUMI meetings and observed the committees during their evaluations on an ongoing basis from the commencement of the study. The Commission did not have enough observers to attend all of the committee sessions, and over the years the number of observers was reduced.

600. Daily notes were made by these observers when they attended an evaluation committee at work (Exhibit R-142), and these notes, which were quite extensive, were entered in evidence during the cross-examination of Durber. Durber had not read the notes himself. He asked Brian Hargadon, one of the Commission's observers, whether there was anything in the observer notes relating to the committee process in particular, which needed to be explored as part of the investigation. Durber testified he received an overview from Hargadon about difficulties in the process of job evaluation, in arriving at consensus, and dealing with the issues. However, at the end of the day, there was nothing in them to be concerned about in terms of the bottom line, that is to say the reliability of the

results. Consequently Durber expressed the opinion he did not consider it necessary for the observer notes to be provided to the Tribunal as evidence in this hearing.

601. Excerpts from the observer notes were read to Durber during his cross-examination and he was asked whether he was given the information

either by Hargadon or in any other context to form his conclusions about the JUMI Study. Some of these excerpts include the following:

> Committee #5
> ...A gender bias problem appears to be developing in this committee.
>
> There is one woman (Sherry) who gives higher scores than the rest of the group for female-dominated jobs and lower scores for male-dominated jobs. She also claims to have first hand knowledge of most jobs and when describing them makes extremely subjective comments reflecting this bias. She will rarely change her rating even if she has taken an extreme position.
>
> There is also a man in the group (Paul) whose ratings reflect the opposite bias. However, his ratings tend to be closer to the consensus rating.
>
> There is another woman in the group (Mary) who, in discussion, appears to have a strong alliance with Sherry. However, Mary's evaluations do not appear to indicate a bias.
>
> Discussion tends to be extremely drawn out in this group as there are consistently opposing views...
>
> (Exhibit R-142, Volume I, page 6)

> Functioning of Committees:
>
> In general, committees have settled into routines which are efficient and also reflect the uniqueness of each group. Given that working conditions are not ideal (i.e. working time is tightly structured and individuals with very different personalities and views must spend extended working hours together) committees are working well.
>
> However, there are a few problems which need to be monitored. I do not know enough about Committee #3 to comment. Committee #5 also has its problems which affect productivity, although not to the same extent as Committee #3.
>
> Members of Committee #5 have problems listening to the views of others. They constantly interrupt each other and often the emotional tenor of the Committee is extremely high.

Committee #5 needs a chair who can be very firm with such disparate and strong personalities. The present chair does not seem to have this capacity...

(Exhibit R-142, Volume I, page 90)

Committee #3

Splits in committee union/management. Job was well written and complete. Louise moved to conform with Jake & Al on K&S. No improvement on committee operations. Atmosphere tense.

Committee #2:

Committee works well.

(Exhibit R-142, Volume II, page 125)

Committee #4:

Took 5 hours to deal with this job (simple). New raters prolonged process, obstinate, even after clarification by consultant.

(Exhibit R-142, Volume II, page 130)

Committee #4:

...Language gender used, Chairperson trying to influence raters.

(Exhibit R-142, Volume II, page 176)

Committee #5:

...Pierre Collard noted a blow-up in Committee 5. He felt it may have been indirectly influenced by the fact that some members of #5 would have no jobs when this process is complete...

Wednesday - the Pay Equity Section, CHRC, rec'd a call from TB to intervene in a blow-out by 2 members of Committee 5.

Thursday - Brian H. and I wandered around committee and things were quiet.

(Exhibit R-142, Volume II, page 203)

Weekly meeting, Monday, October 31, 1988:

132

Ron [Renaud] brought out the point that the ground rules with regard to meeting of the consensus guidelines is not being followed. Result is that after it is all over one party could say that it was not a valid agreement because the rule was not followed, as covered in the procedures guidelines.

(Exhibit R-142, Volume I, page 36)

Additional Observer Notes dated November 24, 1988:

3. Majority vote. Committee 3 & 5 have a problem with this. Apparently they are not following the rules for consensus as spelled out on page 2 of the Working committee Procedures. Committee 2 follow the instructions with no exceptions...

There is also some question on reaching consensus by using the median. Fred had suggested this. For example, you have under working conditions, the following scores, 13, 13, 15, 17, 17. You should settle on 15 as the score. Should this be the solution?"

(Exhibit R-142, Volume I, page 79)

JUMI Committees - Observations:

Today, during my visit to committee number three, I noted that the committee was not observing the two third rule in order to reach consensus. Committee decided to take an average value as a consensus, however, I was consulted in the matter and they went along with my advice. Moreover, a comment was made: "We do not follow this rule unless somebody is here observing us".

(Exhibit R-142, Volume II, p. 102)

Committee #6:

...Also assumption made in working conditions as the committee felt the incumbent was not thorough in filling out the questionnaire.

(Exhibit R-142, Volume II, p. 187)

Notes from Brian Hargadon to Ted Ulch:

I see a couple of problems, at least, with Committee #2.

...

Lack of utilization of the original bench marks. We are told that we as a committee do not have any obligation to follow them. Is this so?

(Exhibit R-142, Volume II, p. 111)

JUMI Committee:

...Keith and Sharon made comments on the analysis they did on their respective committees there was a concern shown by all of the people sitting in for the CHRC that it is obvious there are certain people rating consistently high or low, it may not be resolved soon enough if the information from the tests is not quickly analyzed.

...it has been suggested that we keep out of personal dynamics, that is fingering any person in a committee that may not be up to snuff because it could come back to haunt us. There is a feeling that some committee members, particularly union, are being advised on how to approach the evaluations which would best fit the interests of a specific union membership.

(Exhibit R-142, Volume I, p. 56)

Committee #2:

In position 2317, committee did not follow MEC benchmark and it seems that the position has been overrated...A comment was made:

"It does make a difference to have your presence here during evaluations." People here are not discussing the jobs at all.

(Exhibit R-142, Volume II, p. 180)

Weekly Activities:

...The problem is that committee 5 has well over 100 evaluations to sore thumb and there is a question as to how they were allowed to accumulate so many.

(Exhibit R-142, Volume II, p. 208)

Weekly Meeting, November 8, 1988:

...The members of the committee brought up a number of inconsistencies that have been noted in the various committees. There is a concern as to them being limited to questioning if there is an obvious standard set that may not be followed with other committees.

For example, one committee decided that level D under job knowledge can only be re-asked if the job requires a university degree. Ron asked the consultant if that was the case, and the answer was that was not correct.

Ron will be writing more specifics to be submitted to Ted under separate cover. There is a real concern that these inconsistencies will be allowed to go on and grow in number with the end result that the credibility of the committees, and indeed of us, will be challenged...

Committee #3 is continuing to have some problems. The committee will do their ratings, then search for a benchmark to fit the rating rather than check their rating against an appropriate bench mark.

(Exhibit R-142, Volume I, p. 45)

Consistency JUMI Study:

I would like to bring to your attention what I consider an important issue at this stage of the study and one that should be brought to the attention of JUMI.

Essentially, we should confirm our position that consistency is important; consistency with the MEC discipline and consistency of the five evaluation committees in applying the Willis Plan. I believe we have some legislative authority in the Equal Pay Guidelines in respect to consistency.

...

There have been a number of instances not only mentioned above where committees have for some jobs followed an evaluation process which is inconsistent with MEC and among the various evaluation committees...

-There are other situations like this which makes us concerned about inconsistencies and how we can help ensure that they are corrected as early as possible without compromising our role.

In summary, I recommend that JUMI be advised of our opinion as to how "Acting" situations are to be handled. In addition it would be timely to confirm our position on the importance of consistency; consistency with the MEC discipline and consistency of the five evaluation committees in applying the plan.

(Exhibit R-142, Volume I, p. 47)

Update on Observers Remarks, December 7, 1988:

The observers decided they wanted to go over a number of points that concerned them so a meeting was held this morning.

Before getting into the individual items I want to confirm that we are having some concern shown by various committees during testing time...

The reason for our numbers being diminished at the committees has been discussed with the observers so that we would give the same reason, a) other commitments and b) committees are now requiring less observation because of the time they have been in operation.

Committees 1,2, and 4 are operating quite well. Committee 5 does still have some problems however, they will probably sort themselves out.

Committee 3 is still not functioning up to par. The question arises whether the remaining observers, Sharon and Keith, should spend a disproportionate amount of time in committee 3 because of the problem. So the question remains, do we give preference to committee #3?

When we go back and look at the reason observers from the CHRC were brought into the picture, there is concern that our efforts will be for nothing should a) JUMI fold up, or b) we are to attest to the credibility of both the Master Evaluation Committee and the current five committees in operation.

As it stands now, no observer would attest to the evaluations being fair, balanced and objective. There are too many irregularities within committees and between committees.

(Exhibit R-142, Volume I, pp. 82-87)

602. It should be borne in mind the role of the observers was to act as a "watch dog" in the committee evaluation process. They were to observe, critique and when asked to do so suggest improvement in the functioning of the committees. The observers' notes need to be viewed in this context.

603. Durber accepted the bottom line opinion of the Commission observer, Hargadon, and decided not to rely on the notes as evidence of the reliability of the evaluation results.

604. The Tribunal heard testimony from witnesses who were evaluators on committees and who provided evidence in response to specific observer notes about their particular committee. Having considered Durber's responses to the questions raised during his testimony, the vagueness and lack of specificity of these notes and the responses of the evaluators who testified at this hearing, the Tribunal finds as a fact that the notes do

not significantly impact in a negative sense on the broader issue of reliability.

605. Another aspect of the Commission's investigation involved a three member committee organized by Durber to review re-evaluations conducted by the Treasury Board relating to the Nursing, Home Economics, Occupational and Physical Therapists and Computer Services benchmarks. These re-evaluations are contained in two reports which were presented to the Commission in July, 1990, in response to the Commission investigation into the question of apparent gender bias in the evaluation results. The reports are entitled Evaluation of CS Benchmarks and Corrected Version of NU, Annex B (Exhibit HR-252), and Final Report on Evaluation of Equal Pay Study Questionnaire (Exhibit HR-253).

606. The Commission had asked the Treasury Board whether the Employer subscribed to the observations offered in these reports which raised questions about the specific job evaluations of the multiple evaluation committees. The Commission received no response from the Treasury Board to their enquiries. Durber concluded these reports could be viewed as possible evidence in the investigation, but in the short term, excluded the reports as valid evidence in the Commission's investigation, reserving however, the option to advise the Tribunal of the documents in greater detail. Nonetheless, Durber decided to have a committee explore the substance of the reports (the "Benchmark Review Committee").

607. The Benchmark Review Committee consisted of Esther Brunet, Christine Roberge, an employee with the Commission and Brian Hargadon, an investigator for the Commission. Hargadon and Roberge were trained by Willis. In early September, 1990, the three participants, using the Willis Process, started to re-evaluate each of the evaluations found in the Treasury Board reports. These included 65 benchmark questionnaires. They also examined 203 multiple committee evaluations from the OP, HE, NU and CS Groups. The process, defined by Durber, was that all three committee members had to agree on the evaluation for each job that was re-evaluated. After reaching consensus, the committee then compared their score to the Treasury Board consultant score and the score of the multiple evaluation committees.

608. If the Benchmark Review Committee score was different from the Treasury Board and the evaluation committees' scores, there was an attempt to examine the reason why the scores were different. Then the Benchmark Review Committee gave the benefit of the doubt to the Treasury Board consultants, or to the evaluation committees or failing that, the Committee would justify its own score if it differed from Treasury Board and the evaluation committees.

609. Since Durber was not informed by the Treasury Board as to the purpose of the reports provided in July, 1990, his conclusions were

primarily based on the conclusions contained in the Benchmark Review Committee's report.

610. Brunet did not participate in writing the Committee's final report (Exhibit

HR-254). It was prepared by the Commission members, Roberge and Hargadon and was reviewed by Durber. The conclusion contained within the report and attested to by Durber is that no weight should be placed on the Treasury Board reports. The Benchmark Review Committee's examination confirmed the JUMI evaluations with very few exceptions.

611. An earlier draft of Exhibit HR-254 was prepared by the two Commission members of the Committee, and is dated June of 1991. That draft was introduced in the cross-examination of Durber as Exhibit R-140. There were two passages, at pp. 26-27, which were not included in the final report. These pages refer to sore-thumbing and difficulties experienced by the evaluation committees in the use of benchmarks. Durber removed these from the final report. It is his opinion, these pages were not particularly "relevant to what they [the Benchmark Review Committee] were doing..." (Volume 159, p. 19790). Durber instructed these pages be dropped from the final version. In his view they were interesting comments on difficulties encountered with benchmarks but did not add to what the Commission already knew. In his opinion, they were more instructional for use in future pay equity exercises.

612. Durber testified he asked both Roberge and Hargadon about the considerations raised on pages 26 and 27 of the original report, Exhibit R-140. Durber testified he was told that the purpose of these two pages was to comment upon "lessons learned, and their own perceptions of the difficulties the Commission might encounter in fulfilling their observer role in future initiatives." The Commission would, as a result, be forewarned of the problems which occurred during the JUMI Study including the difficulties with the rationales. Durber did not consider their comments as solid evidence, but more as useful material for future work of the Commission.

613. With respect to the report of the Benchmark Review Committee, Durber considered that the matters contained in pages 26 and 27 would come forward through Willis during the Tribunal hearings. Durber claimed the Commission had neither the resources nor the time to begin an investigation of the MEC process while preparing for its participation in these hearings.

614. The Tribunal did have the benefit of Brunet's testimony relating to pages 26 and 27 of Exhibit R-140, which appears in Volume 214, at p. 27852, lines 5 - 15:

> I noticed that pages 26 and 27 made me smile when I saw them because, when I was working with Christine and Brian Hargadon, Jim Sadler was heading the study from the Northwest Territories. He would often come and see how things were going, and all that. Once we found out that he was going up there, we said, "How about we share some information that we have, so that you can bring it up.
>
> When I saw pages 26 and 27, a lot of that I had input in.

615. Brunet was under the impression she would be called upon to review and sign the report. In fact, she was not asked to do so but she

did, however, receive a copy of the report. She testified while the committee was doing its work, Jim Sadler, an employee of the Commission who was heading a pay equity study in the N.W.T., often came to see how the Benchmark Review Committee was functioning. The Benchmark Review Committee suggested that they share information with Sadler so he could take it with him to the N.W.T. study.

616. Both Brunet's understanding of the comments contained on pages 26 and 27, and Durber's opinion as to their usefulness, are corroborated in Exhibit R-141, a letter written by Sadler addressed to a union representative involved in a pay equity study in the N.W.T. This study is referred to as the Joint Equal Pay Study (JEPS) which was using a newer version of the Willis Plan. Some of Sadler's comments in that letter were based on discussions he had with members of the Commission's Committee. Those discussions corroborate both Durber's and Brunet's evidence about the Committee's perception about sharing this information with the Commission.

617. Durber's opinions and conclusions about Exhibits R-140 (Draft Report) and R-142 (Observer Notes) led him to decide not to introduce these documents as part of the Commission's case. The Tribunal hearing is in the nature of a public enquiry, and the Commission's role is to represent the public interest. Decisions about the relevance of documentation garnered by the Commission during its investigation of the s. 11 complaint is within the purview of the Commission. In circumstances such as these however, the Commission's decision to exclude these documents from its case is open to

criticism if the documents are found to be relevant and sensitive to the issue of reliability.

618. Before proceeding further, the Tribunal is of the view the reports in question, namely exhibits R-140 and R-142, should have been introduced in their entirety as part of the Commission's case with accompanying explanations. The decision as to their usefulness ought to have been left with the Tribunal. The Commission's case would have been better served had the entire exhibit been entered in the first place.

619. During cross-examination Durber offered a further explanation as to his reasons for not interviewing Wisner. He conducted an ex post facto review of Wisner's rationales for purposes of what he described as clarification. Durber used both the committee's and Wisner's rationales to do this analysis. It involved a review of each difference and a determination of the extent these differences cancelled one another out. After Durber categorized the differences between the committee and consultant, he looked at the numbers to determine whether the distribution of these differences were patterned or random.

620. Willis was asked to comment on Durber's analysis which was based on the examination of rationales. Willis replied in Volume 208, at p. 26939, he had trouble with Durber's conclusions. Willis doubts very much if bias could be recognized by looking at rationales. In Willis' opinion, bias is very subtle and not something that can be looked at on a job by job basis. Willis testified in Volume 208, at p. 26939, lines 8 to 13:

> You have to look at a total pattern and, to me, it would be totally inappropriate to single out certain ones of those re-evaluations and say, "We will discount those." I think you either take them all and look at them at their face value or you don't take any of them.

621. According to Willis, if his consultants are doing an evaluation during the course of the study, the reasons for the differences are very important as they will provide the consultants with some basis for retraining of a committee. Willis recognizes there is always going to be some random variance, and random disparity after the study is completed, and therefore, he does not, at this stage, concern himself with the reasons. In the context of Durber's analysis, Willis said he always expects some differences between consultants and committees, but he did not see any value in attempting to use those differences to analyze whether or

not there is a problem. Willis elaborates further in Volume 208, at p. 26944, lines 17 - 23:

> A. What I have said or at least what I intended was that since bias is a very subtle thing, I think our only opportunity for examining the extent to which there is a different interpretation of male versus female jobs is by looking at the total results after the study has been completed.

622. The analysis done by Durber was presented in mathematical form as numbers and tables and conclusions about symmetry between numbers and whether these numbers were demonstrative of patterns. The Tribunal's view is that this analysis has a statistical component because of the particular methodology used by Durber. Without the assistance of a qualified statistical expert, we are unable to properly interpret Durber's analysis which, therefore, must be disregarded.

623. In 1992, during the appearance of Willis before the Tribunal, Durber decided to further investigate the quality of job information contained in the questionnaires. Accordingly, he retained a researcher, who had no experience in job evaluation but who had "pretty good analytical ability" for the purpose of examining a cross-section of the questionnaires. The cross-section included 63 benchmarks, 587 non-benchmarks for a total of 650 questionnaires. The researcher did not appear before the Tribunal.

624. The researcher's task was to look at the information to assess completeness, consistency, legibility, and whether the safeguards had been followed and finally, to determine if there was an indication each questionnaire had been validated by the employer's supervisor.

625. Durber's evidence is he discussed with the researcher some of the characteristics that could lead to deciding whether or not the questionnaires were complete. In this regard, Durber prepared some procedures and questions for the researcher. As background, the researcher was provided with the purpose of the job information, the process used during the study to collect and screen the information, as well as

information for identifying basic data such as department, questionnaire number, occupational group and other such information.

626. This project took the Researcher two months to complete. A meeting between Durber and the researcher occurred every week to discuss

problems. Durber personally reviewed any questionnaires where problems were encountered, which involved approximately 5 per cent of the questionnaires. Durber testified he closely supervised the researcher during the examination of benchmark questionnaires.

627. The following is a list of criteria used by the researcher in this exercise:

1. Legibility - can the questionnaire be read?
2. Language - whether the questionnaire was French or English?
3. Script - whether the questionnaire was typed or hand written?
4. Signature - whether it was signed or not?
5. Comments - whether the supervisor commented?
6. Completion - whether all of the parts of the questionnaire had been completed?
7. Consistency - whether supervisor was consistent with the incumbent?
8. Notes - whether there was evidence of interviewer or reviewer notes?
9. Facts - whether the questionnaire contained fact versus editorial comment?

628. The report entitled An Examination of the Quality of Questionnaire Information used by the Federal Pay Equity Study (Exhibit HR-245), contained both findings and conclusions about the completeness and accuracy of the job information. In the Tribunal's view, Durber is expressing, in the report and in oral evidence, the opinions of his researcher which may or may not be well founded. Due to the researcher's lack of expertise in pay equity job evaluation, it is the Tribunal's conclusion it must reject any opinions contained in this report. There is, however, factual content in the report, not based on opinion, which in our view is helpful. These are listed as follows:

Findings:

¨Required questions were answered 95% of the time.

¨Supervisors provided signatures on just over 99% of questionnaires. In just over 96%, the supervisor commented, seeming to contradict incumbents about 9% of the time. In 95% of these contradictions, subsequent interviews clarified the work.

¨In two-thirds of the files, interviews were carried out, with supplementary information provided. The investigator noted that the latter was frequently extensive...

¨Legibility of the description in questionnaires was in all cases good.

Conclusions:

¨There was a system for reviewing and assuring the completeness of the information about work in the Joint Initiative.

¨There was a system for ensuring the accuracy of the job information...through supervisory review.

¨Those involved in reading questionnaires made efforts...to obtain further information to improve their understanding...where the supervisor and incumbent appeared to disagree about the work. (Exhibit HR-245)

(ii).Sunter's Analysis

629. The Commission asked a former director of Statistics Canada, Alan Sunter, to examine the full set of data from the Wisner 222 and the Willis 300 and look for patterns relating to gender composition. The Commission also requested Sunter, to assess the statistical significance of the formulae relating to possible gender bias used by the Treasury Board in its March, 1990 methodology paper.

630. Sunter, a qualified statistical expert, did not have a background knowledge in pay equity prior to his involvement with the JUMI Study results. He became involved in the analysis of the JUMI data as a result of a request by Durber on April 6, 1990 who asked him to attend the workshop scheduled for April 9, 1990. The workshop was to focus on the Treasury Board methodology document (Exhibit HR-185). Sunter testified he was unable to contribute in a constructive way to the workshop and he simply listened to the discussions. After the workshop, he met with Durber and began to realize there had been a large study addressing the question of pay equity between male- and female-dominated occupational groups. He also learned there had been subsequent re-evaluations of samples taken from the evaluations. This led to the question of whether there was gender bias in the evaluations. This was a matter of concern to the Commission.

631. The statistical evidence concerning the question of gender bias in the evaluation results was provided by Sunter and Shillington, both experts in statistics. Shillington was not employed by the Commission to do any statistical analysis of the results. However, because of Shillington's involvement in the IRR testing and other aspects of the JUMI

Study, he testified before the Tribunal. During his appearance, he was requested to provide opinions on Sunter's analysis.

632. Sunter was asked specifically by Durber to perform three analyses. Firstly, he was to look at the question of gender bias in the re-evaluations and for this purpose, he was given two sets of data, the Wisner 222 re-evaluations and the Willis 300 re-evaluations. Secondly, he was provided with the whole data set from the JUMI Study and was asked to examine the question of equal pay for work of equal value between male- and female-dominated occupational groups. Thirdly, he was given the Treasury Board methodology document (Exhibit HR-185) and asked to examine

specifically the Treasury Board methodology and offer whatever criticism seemed appropriate.

633. Sunter's interpretation of the term "gender bias" used in his analysis of the data is provided in Volume 102, at p. 12275, lines 3 - 17:

> A. I supposed gender bias to mean that there would be some systematic tendency of the evaluation committees to underscore positions from male-dominated occupations or to overscore positions from female-dominated occupations or perhaps both of those things.
>
> Q. What do you mean by "systematic tendency"?
>
> A. At this point, of course, I didn't know, but since the term bias had been used, then I assumed that bias would have to mean a consistent tendency that would display itself in some kind of recognizable pattern in the data, that I would see that when I looked at the data and performed some kind of analysis on the data.

634. Willis testified a consultant trained and experienced in the application of the evaluation system possessing an objective view point, can be expected to evaluate consistently and without a predilection towards either male or female dominated jobs, or towards either management or union sides. Willis asserts consultant evaluations are useful in examining the consistency of committee evaluations, and, more importantly, in assessing any pattern of bias which may have occurred. Willis' view is that his consultants' experience, background, intent and philosophy has always been not to favour one side or the other but to walk the middle road. Willis' objective in doing the re-evaluations was to identify whether or not there

was a gender based pattern or difference in treatment between male- and female-dominated jobs. Willis referred to the differences between consultant and committee as disparities. It is within this framework Sunter began to examine the data.

635. Sunter testified statisticians collect and analyze data with two quite distinct concepts. One he refers to as "descriptive" and the other as "analytic". In his view, the distinction between these two broad areas of enquiry is important in respect of the work he did and his interpretation of the JUMI data.

636. Sunter compared the two re-evaluation data sets, the Wisner 222 and the Willis 300 against the committee evaluations of the same jobs to see whether statistically there was a patterned difference in the manner in which evaluators treated different types of positions and if so to measure the size of the differences he found.

637. Sunter performed a statistical test known as a t-test to measure whether there was a difference between the treatment of male and female questionnaires by consultants and committees using only the Wisner 222, then using only the Willis 300 and then pooling the two data sets together.

638. According to Shillington, who also performed t-tests in his IRR analysis, the t-test is a statistical test that summarizes information about how far two averages are from each other. In this case, the statistician is looking at the male average and the female average to see if there is evidence they are treating male and female questionnaires differently. He states in Volume 86, at p. 10668, the t-test hinges on three things:

1. How far apart are the two averages? The more apart the two averages are, the more likely it is to say yes it comes from different populations; yes, this person is treating male and female questionnaires differently.

2. The larger the sample size the more likely it is to say that there is significant evidence that they are treating the two populations differently.

3. The more concentrated the values are, the easier it is to say that this is a true pattern.

639. If the difference in the average scores is substantial then, according to Shillington, it is more likely you will get a "significant" result in statistical terms of measurement. A significant difference reflects a true difference between two groups and will demonstrate the result, most likely, could not have happened by chance. Statistical significance in this context pertains to mathematical probabilities and whether the numbers are unlikely to have happened by chance. (Volume 87, p. 10673).

640. Sunter testified about the limitations of the t-test. One such limitation is that when the sample is very large, even if the difference is minuscule, the t-test would find it to be significant. In other words, the t-test rejects the null hypothesis of no difference when the sample is large enough. Another limitation is that the t-test is not attentive to differences of practical importance, it simply follows a mathematical routine of testing the null hypothesis of no difference against the alternative hypothesis that a difference exists.

641. Sunter found the size of the difference in the treatment of positions from male- and female-dominated occupational groups by committees and consultants was 2.3 per cent in the pooled data. He performed further t-tests to determine if the consultants and the committees differed in their treatment of female-dominated positions. The results showed that for positions from female-dominated jobs, there was no statistically significant difference between the manner in which the consultants and the committees rated these positions. For positions from female-dominated occupational groups, the consultant and committee ratings are not significantly different whether one compares the committees to the Wisner 222, the Willis 300 or the pooled consultant re-evaluations (522). The size of the non-significant difference in the treatment of positions in female-dominated positions for the pooled data was 0.05 per cent. For the Wisner 222, this difference was 0.02 per cent and for the Willis 300, this difference was 0.07 per cent (Exhibit HR-191).

642. Sunter then performed the same t-test on the male-dominated positions. He determined that the consultant and committee ratings were significantly different for positions from male-dominated occupational groups. The size of the difference between the committee and the consultant treatment of positions from male-dominated occupational groups depended on which of the consultant re-evaluations, the Wisner 222 or the Willis 300, were used as a basis for comparison with the committee results. It also depended on whether the committee or the consultants were placed in the denominator of the equation. Sunter testified since there is no "true

value" for any given questionnaire, there has to be some standard by which to compare committee and consultant evaluations. When it is contended the committee is biased relative to the consultant, Sunter states the consultant is taken as the baseline or standard of comparison and the consultant scores are found in the denominator of the equation to determine any difference in treatment.

643. The size of the difference in the treatment of male-dominated positions for the pooled consultant re-evaluations (522) was found to be 1.8 per cent, when the consultant evaluations are used as the denominator. For the Wisner 222, this difference was 2.5 per cent and for the Willis 300, it was 1.3 per cent (Exhibit HR-191).

644. Having found the consultant and committee ratings were significantly different for positions from male-dominated occupational groups, he testified the size of the difference in the treatment of male-dominated positions was twice as great in the Wisner 222, a difference of 2.5 per cent than in the Willis 300, a difference of 1.3 per cent.

645. Sunter preferred using the Wisner and Willis pooled result (522) as more reliable in establishing the size of the difference between the committee and the consultants rather than using either the Wisner 222 or the Willis 300 independently. This difference is stated as 2.3 per cent.

646. As to whether there was any pattern in the differences between the committees and the consultants, Sunter found in over half of the evaluations between the consultants and the committees there was no difference at all. In separating the data, he found in about one-third of the comparisons between the Wisner 222 and the committees there was no difference and in about two-thirds of the comparisons between the Willis 300 and the committees there was no difference. He found it inconceivable that, given this number of agreements, there was a consistent pattern of discrimination.

647. Sunter testified having found differences between the committee and the consultant in the treatment of male questionnaires, he would not conclude the committee was biased or that the consultant was biased. In his opinion, the only conclusion to draw was that both the committee and the consultant appear to have a bias relative to each other with respect to male evaluations. Sunter went on to say you may call this a relative bias, or you may attach the term gender bias to it. However, he had difficulty with the term "gender bias" because without further testing, one could not conclude whose gender bias it is and whether the bias is merely incidental

to gender or whether it is contingent on something else, which itself is incidental to gender.

648. The crucial question at this juncture in Sunter's evidence is whether the t-test results indicate a systematic pattern in the disparities or whether the differences are merely random. The Commission submits systematic patterns of gender differences must, by definition, be differences which are demonstrative of a system at work, something regular or methodical. (Para. 199 of written submissions).

649. The Employer submits a different treatment of male and female questionnaires is indicated by a pattern in the disparities such that the evaluation of female jobs systematically differ from the evaluation of male jobs. (Para. 289 of written submissions). The Employer's interpretation of pattern can be better understood in the following exchange with Sunter which appears in Volume 217, at p. 28243, line 8 to p. 28244, line 1:

> Q. Mr. Sunter, I am just talking about the chi-square and the T test when you split the questionnaires by male and female. There was a pattern there.
>
> A. There is a difference in the pattern. I wouldn't use the term "pattern". There is a difference. We have acknowledged the difference. We are trying to explain the difference.
>
> Q. But there is a difference in treatment, let's put it that way.
>
> A. There is a difference in the average -- I don't like the term "treatment", I must say, because it implies some physical process. There is a difference in the differences between consultant and committee scores. You may use the word "treatment" for that if you would like, but I prefer not to use the word "treatment".

650. Sunter then attempted to explain and understand the differences between committees and consultants by fitting models to the data which he says are necessary in order to attach meaning to the "notion of gender bias". It is in this area of his analysis where Sunter emphasizes the distinction between the descriptive use of statistics as opposed to the analytic use. The latter use involves his adaptation of models to data. Sunter testified if gender bias is present in the results, a statistician expects to see some degree of consistency across evaluations which are somehow related to gender. Therefore, he tested the data for consistency by using models to illustrate how gender bias might operate.

651. Sunter examined three plausible models to explain how gender bias might affect the committee's results. For example, one such model he termed "additive" which he described as a constant addition by the committee to the consultant scores or a constant subtraction by the consultant from the committee scores. Sunter eventually disposed of all of these models because the data did not support such configurations.

652. Sunter again tested the differences between the committee and the consultant by using the chi square tests. He also applied this test to the Wisner 222, the Willis 300 and the pooled data. All of these tests indicated statistically significant results. Sunter criticized the usefulness of chi-square analysis in these circumstances. In his opinion, the chi square tests are not helpful in understanding the difference between the treatment of male- and female-dominated jobs by the consultants and the committees. His concern about the chi square test is that this test measures the frequency rather than the size of the difference, as is the case with the t-test. Therefore significant results from the chi-square test can be misleading about the real difference between the numbers. Accordingly, he preferred to use t-tests which showed a difference of 2.3 per cent from the pooled data as best representing the size of the difference between committees and consultants.

653. Having seen no difference between the consultants and the committees, on average, for female-dominated occupations, Sunter went on to explore the idea of gender bias being an unconscious discrimination for or against occupational groups by gender. He suggested, that the way gender bias might work in this context, is that there are certain underlying male characteristics or female characteristics and that occupational groups that have more males will tend to show this pattern of discrimination rather strongly. Having tested for that, he did not find any such correlation between the degree of maleness of an occupation and a pattern of relative differences. Sunter concluded from his analysis that he was unable to find any consistent pattern of differences, and that there was no plausible or conclusive explanation for the differences between the committees and the consultants.

654. Sunter concluded from his analysis that without a level of consistency in the incidence of differences along gender differentiated lines between committees and consultants he was unable to conclude the difference was attributable to gender bias. He says in Volume 102, at p. 12277, line 25 to p. 12279, line 1:

A. My general conclusion on the question of gender bias -- mind

you, I still don't know what gender bias is, you understand, but my general conclusion on this was as follows. There was a slight difference between -- there was virtually no difference between the consultants and the committee on positions from female-dominated occupations. This could be put aside.

On positions from male-dominated occupations, there is indeed a difference, not large but indeed a statistically significant difference, between committee evaluations and consultant evaluations. This does not lead me to the conclusion, however, that there is gender bias, putting aside for the moment that I still don't quite know what I mean by gender bias because there are other possible explanations...

...

147

A. In order to conclude that this was gender bias, I would have to find some kind of consistency in the observations. I was unable to find the kind of consistency that would enable me to reach that conclusion.

655. He also found the lack of consistency in the differences and the absence of an alternative plausible model of gender bias did not justify adjusting committee scores in the manner adopted by Treasury Board in their 1990 methodology paper.

656. Sunter returned to the question of gender bias and explored other factors which occurred to him and were not pursued in his initial investigation. His exploration was with factors that might be associated in some way with gender and, therefore, considered possible causes for the difference in the scores other than gender bias. He examined other characteristics, such as perceived salary, nature of work, size of group, which he thought might be correlated with gender. Simply expressed, in ordinary language, Sunter explored the degree of association between the differences and some of the other characteristics of the data.

657. One characteristic which Sunter noted between male and female questionnaires is that the data showed female questionnaires coming from a small number of relatively large occupational groups. Male questionnaires, on the other hand, were coming from a large number of relatively small occupational groups. Sunter postulated evaluators might be more familiar with the female-dominated occupations which included jobs such as clerks, secretaries and nurses, than with the male-dominated occupations which

included air traffic controllers, defence research scientists, patent examiners, etc. He divided the databases according to size of the group, and using group as a proxy for familiarity with the type of work, he compared the differences between the consultants and the committees for the Wisner 222 and the Willis 300 data. Although the results of this statistical analysis did not indicate statistically significant results, Sunter believes they did demonstrate a strong association between size of group and the pattern of differences between committee and consultants.

658. Another characteristic he noted which differentiated between male and female questionnaires, is the relative distribution of positions from male- and female- dominated occupational groups across the range of evaluation points. He found 75 per cent of questionnaires from female-dominated occupational groups fell below a certain point value while only 25 per cent of positions from male-dominated occupational groups fell below the same value. He hypothesized any bias that relates to point distribution, such as a bias in favour of placement in the hierarchy of jobs, or bias in favour of or against managerial or supervisory positions, or a bias in favour of the skills acquired in post-secondary education, could look like a gender bias.

659. Sunter then performed several comparisons to see if the differences between committee evaluations and consultant re-evaluations were associated with the relative distribution of questionnaires in the high or low point range. Again, in this comparison, the results did not demonstrate a statistically significant difference. He concluded, however,

they did show an association between high and low points and the differences between committees and consultants when split along gender lines. Sunter referred to this bias as a "point bias or value bias", that is, the higher the value of the job the more likely there is to be a difference between the committee-assigned score and the consultant-assigned score.

660. Willis responded to Sunter's evidence about value bias during his second appearance before the Tribunal, which followed Sunter's testimony. Willis said he would like to see further analysis as to whether the differences between the committees and the consultants might be associated with value bias. Willis wanted to know if 10 per cent of the high evaluation scores were removed from the database, whether the extent of the differences between the consultants and the committees would be reduced. On this point, Willis says in Volume 211, at p. 27491, line 19 to p. 27492, line 4:

I had said I would rely on a statistician. This task was not given to me, but if it had been given to me and my statistician had said there is an appearance of bias here and it doesn't necessarily represent bias, I would say "Okay, let's take those top ones out and let's see what it looks like then." Maybe it will be less than 1.8 per cent and maybe it won't. Since we are dealing with several million dollars, my suggestion would be that if it doesn't change that percentage, then I would tend to adjust.

661. As a result of Willis' comments, Sunter performed an additional analysis to determine whether the differences between the consultants and the committees could be reduced by "value effect". His analysis, which is termed value effect, was introduced by the Commission in response to the question raised by Willis. Sunter defined value effect in Volume 216, at p. 28049, line 23 to p. 28050, line 1:

A. The value effect would be some systematic tendency for differences between consultant and committee to show up in association with increases in value of the job.

662. Sunter's further statistical work explored how much of the difference between committee evaluations and consultant re-evaluations could be attributable to "value bias". By this he meant the difference between how the committees and the consultants treated high and low point questionnaires. Sunter's analysis included statistical methods for standardizing the data because of what he described as a distribution problem. Because of this problem, he could not merely discard 10 or 20 per cent of the top end scores as suggested by Willis. On the basis of this analysis, Sunter concluded that at least one half of the apparent gender differences between the committees and the consultants is immediately accounted for in differences in value distribution.

663. Relying on the analysis he performed (Exhibit HR-265), Sunter testified, once he removed the value effect, the overall difference of 2.3 per cent between the consultants and the committees was reduced by 1.2 per cent.

664. There has been doubt expressed by Shillington, on whether or not statistically or otherwise, you can separate two data analysis issues, one being whether or not a pattern is related to gender, and the second being whether or not the pattern is related to the scores being high or low. Shillington explains this problem in Volume 131, at p. 16045, line 23 to p. 16046, line 15:

The regressions were done in a way to try to see if there was a relationship between the differences between the consultants and the committee in gender.

It is also possible that any differences that might have existed between the consultant and the committee scores were not directly related to gender but perhaps were related to high values versus low values. This has been talked about here.

The confounding is introduced because there is a strong trend in the data for the male questionnaires to all have high values relative to the female and the female questionnaires have a fair tendency to come from the lower end of the spectrum, which means you cannot separate those two data analysis questions, or it is difficult to separate them.

And also in Volume 131, at p. 16048, line 16 to p. 16049, line 11:

In this circumstance, back to the analysis of the Willis scores and the possible adjustment, we have a situation which -- to the extent that there is a pattern here, if someone came and said this is possibly not due to gender, maleness or femaleness, but rather could be due to professionalization or some questionnaires having much higher values than others, you would have a problem extracting those two separate hypotheses from the analysis because you have a situation in which the males predominantly had high values, the females predominantly had low values. So maleness is confounded with high and low values.

That is reflected in the distribution. That is why it is a distribution question. The distribution of the Willis scores for the males tended to be quite a bit higher than the distribution of the Willis scores for the females. It is a confounding issue. That is why in interpreting it you are going to have to be cautious about that.

And further on this point, he says in Volume 131, at p. 16051, line 12 to p. 16052, line 5:

THE CHAIRPERSON: ...But just looking at these and what you can say about what they describe in terms of their distribution, what you can interpret from that is that the males tend to be high, the females tend to be low, but you can't, because of this confounding effect, you can't really interpret anything else with certainty. Is that ---

THE WITNESS: That is right. You have to be very careful when interpreting the results because you have to keep in mind that if somebody came with an alternative explanation for the data and the explanation was that this had nothing to do with gender, that this was high score/low score effects, you have collected your data in such a way that most of the high scores are males and most of the low scores are females. So they are two equally valid explanations for the same data.

665. While Sunter acknowledged difficulties in unconfounding data, he said he was able to isolate or distinguish from the disparities, a portion that could be attributed to different value distributions of the male and female questionnaires. Sunter maintained he did not find it difficult to make a differentiation between gender and value and he could unconfound the data to this extent. Under cross-examination by Respondent Counsel, he was not prepared to agree that gender is a proxy for value or that value is a proxy for gender. He did agree, however, there are many factors correlated with gender, and if the difference between committees and consultants stems from some other causal factor, which itself is associated with gender, then he could never determine how much of the difference would be attributable to gender bias. (Volume 217, p. 28247).

666. Sunter believes the question of association of the differences in scores with other characteristics in the data becomes important if there is going to be some adjustment in the committee results to eliminate gender bias. In this context, Sunter believes it is important to demonstrate the magnitude of gender bias, how it operates and how it can be adjusted out of the actual data. Sunter believes the association of the differences in scores with value bias becomes vital at this stage.

667. Sunter concludes the whole question of association with other characteristics is intimately connected with the process of adjustment. Accordingly, Sunter found it difficult to separate the question of how to analyze the data from the question of what you wish to do with the results.

668. Sunter was aware of the Treasury Board's methodology paper in which the Treasury Board used and adjusted the Wisner 222 data when calculating the equalization payments of January 1990. Sunter refers to this adjustment as an "across-the-board" adjustment. He describes what he means by an across-the-board adjustment of evaluation scores in Volume 103, at p. 12426, lines 16 - 22:

What I do, if I am about to make an across-the-board adjustment, let us say, of values assigned to questionnaires from male-dominated occupations, would be to say, "Let us increase all of these, all of them, by four per cent without exception." That is what I mean by an across-the-board adjustment.

669. In Sunter's view an across-the-board adjustment requires some consistency in the pattern of gender bias, and an across-the-board adjustment can only be made on the basis of an across-the-board bias. He explains this in Volume 103, at p. 12427, lines 8 - 10:

...these are two sides of the same coin. If I cannot find the one, it seems to me that I cannot be justified in doing the other.

670. According to Sunter, the Employer performed a regression analysis, another form of statistical measure, on the Wisner 222 data as described in their methodology paper (Exhibit HR-185). The regression analysis conducted by the Employer assessed differences in treatment between committees and the Wisner 222. The regression analysis was the basis upon which the Employer calculated the unilateral adjustments to the scores in January, 1990. A critique of the Treasury Board's approach, given by Sunter, included an analysis of "overlapping confidence regions" of regression lines that represented scores for male- and female-dominated jobs.

671. It was Sunter's opinion the Treasury Board's regressions should not have been used to adjust the scores from the female-dominated occupational groups at all. With respect to the male data, the regression line comparing the Wisner 222 re-evaluations and the committee scores were significantly different over the second half of the point range of scores. Sunter found that the overlap of the male and female confidence regions, up to the 250 Willis point mark, is not strong evidence the consultants and the committees differed significantly or consistently below 250 Willis points.

672. Sunter concluded from his analysis of the regression lines there appeared to be no difference between the consultants and the committees for at least three-quarters of the female questionnaires. Accordingly, he found no justification in the Treasury Board regression lines for making relative adjustments to all of the male and female questionnaires.

673. Shillington, under cross-examination by Respondent Counsel, indicated he did not have any problems with the way Sunter conducted his

analysis of the Treasury Board's adjustment methodology. He was of the opinion Sunter had drawn a reasonable conclusion from his analysis. (Volume 136, pp. 16741-42).

674. The Tribunal did not hear any expert evidence concerning Treasury Board's methodology of adjusting scores, other than what has been was provided by Sunter and Shillington about their understanding of the methodology contained in Exhibit HR-185.

675. Sunter testified the use of regression analysis to identify differences in evaluation scores between Wisner and the committees, is an unsuitable statistical tool. The regression equations, in his estimation, do not provide support for the Treasury Board's adjustment of female questionnaire scores downward, which average 3 per cent overall and male questionnaire scores upward, which average 4 per cent overall. In Sunter's opinion, which is supported by Exhibit HR-213, the regressions predict for the first three-quarters of the female questionnaires, either an increase in the female questionnaire scores or no change at all.

676. Insofar as the three areas Sunter was asked to review at the request of the Commission, his conclusion on the gender bias analysis in

the first two areas are: (i) there was no where near the level of consistency in the incidence of differences along gender differentiated lines which would enable him to conclude there was gender bias. Sunter testified this is not to say there is not gender bias, only that one cannot conclude there is gender bias and a review based on that finding of the Treasury Board methodology leads him to conclude there is no basis on which the Treasury Board could have justified any adjustment of the committee scores. The third aspect which deals with an analysis of the differences in compensation from male- and female-dominated occupational groups, is not in issue at this stage of our decision.

F. ROLE OF CONSULTANTS IN RE-EVALUATIONS

677. Both statistical experts testified under cross-examination consultant scores can be used as a reference point to compare committee and consultant scores on the assumption the consultant scores are free of "gender-related bias". This is a term introduced by Respondent Counsel to describe a bias unrelated to gender but to some other characteristic which is itself related to gender.

678. Both statistical experts expressed the opinion they preferred committee scores over consultant scores. Shillington, in particular, found it difficult to accept that any individual could be free of gender related bias and he says the following in Volume 139, at p. 17084, line 4 to p. 17085, line 2:

> A. I think that is more in the line of a decision that could be made. You have indicated that the issue of gender-related bias is the area of concern and not being as concerned as to whether or not it was directly related to gender or not. So I think deciding not to be concerned with the reason that the gender-related bias, if there is evidence of that, is present -- that is a decision.
>
> If that sentence is to be interpreted to mean "if you decide that you don't care for the reason, then you don't need to look for it", you are right. But I certainly never -- several times in testimony you asked me to assume that Mr. Wisner was without gender-related bias and I more than once said "How can that be. How can someone be so free of thoughts about high score/low score, dirty work/clean work. How could this person be equally familiar with all jobs", but you asked me to assume that.
>
> So, I am not sure that the sentence the way it is presented there is a fair or complete summary of my opinion about this, and I certainly can't speak for Mr. Sunter.

679. The position of the Employer essentially is the consultants' re-evaluations are only used in the statistical analysis as a point of reference for determining whether there is a pattern of different treatment of male and female questionnaires by the committees. Willis testified the consultant scores are not to be substituted for committee scores, therefore the Employer submits using the consultants' re-evaluations as a reference point does not mean the consultant re-evaluations are to be preferred to

the committees', because there is no substitution of scores. However, the Employer contends, for purposes of using consultant re-evaluation to determine a pattern of different treatment, the Tribunal may prefer the consultants' relative treatment of male and female questionnaires without preferring their scores on any one questionnaire. (Respondent's written submissions - paras. 319 and 320).

680. Shillington expressed the opinion in using the consultant scores as a reference point, an assumption had to be made the consultant scores

were to be preferred to the committees'. He gives the following response in Volume 136, at p. 16692, line 16 to p. 16693, line 15:

> Q. When we are using the consultants as a reference point only, we are not saying that we prefer the consultant's score on any one questionnaire over the committee score. We are only making the assumption that the consultant scores across the board are free from gender-related bias.
>
> A. But that you are not preferring them?
>
> Q. But that we are not preferring them. So, we won't take the score on any one questionnaire and say the consultant scores are better. That's not a necessary assumption.
>
> A. But I still think you have to end up assuming they are better and the example again is when I used -- suppose that the consultant didn't look at the questionnaires at all and the consultants just wrote down daytime temperatures, blood pressure, whatever. Right? They would certainly not be preferred and they certainly would not exhibit a gender preference if they just ignored the questionnaires totally. So, I think you do have to assume that the consultant scores are to be preferred.

681. Sunter testified the committees should be preferred to the consultant's for four reasons. His first reason is based on his own experience in the field of statistics which led him to conclude committees often apply a system better than the consultant who developed it. His remaining three reasons for supporting committee evaluations over consultant evaluations are based on his analysis of the data. One of his analysis tested for consistency between Wisner and the Gang of Four.

682. Sunter tested for consistency between Wisner and the Gang of Four by performing statistical tests such as t-tests and chi square analysis. The results he obtained confirmed, in his mind, that Wisner and the Gang of Four, differed among themselves. Sunter's conclusion was that if the consultants cannot agree among themselves, it cannot be the case that the consultant is always right. His analysis led him to conclude the consultants were not consistent among themselves and on this basis the committees should be preferred.

683. When Sunter was called as a reply witness by the Commission in November, 1994, he testified he had undertaken a further analysis on the question of the relative reliability of the committees and the consultants.

Sunter also used standard statistical measures in the form of regression analysis, to support the use of the committees as a point of reference in any analysis of gender bias. Sunter formed regression line comparisons using two sets of the data, the MEC scores and all the scores on which the committees and the consultants agreed, which led him to conclude any notion of committee bias for male job evaluations could not be sustained.

684. With the exception of Sunter's further analysis given in reply evidence, the remainder of Sunter's analyses were commented on by Shillington. Shillington concurred with Sunter's statistical conclusions, with the exception of one analysis, namely, Sunter's variance co-variance analysis. Shillington had an opportunity to meet with Sunter to discuss this analysis. Having had that opportunity, Shillington continued to maintain he had problems with drawing the conclusion from the variance co-variance analysis that the consultant is to be preferred to the committee. Dr. Shillington offers the following explanation in Volume 133, at p. 16306, lines 8 - 22:

> So, I would have a difficult time believing that the data can help you unravel that that the data can actually help you decide that one rater is preferable to the other, unless you had a third set of numbers which you believe to be the correct values.
>
> So, I look at the models and I say the models look reasonable and, yes, it's clear that the correlation matrix in one case is closer to the observed data than the correlation matrix in the other case, but even after discussing this, I have to step back and say: This may be true, but how can the data help you unravel which rater is better if you have no third set of numbers, which is the correct values?

685. Shillington went on to say his opinion on this aspect of Sunter's testimony did not distract from his approval of Sunter's analysis on the issue of gender bias. He responds as follows in Volume 133, at p. 16306, line 23 to p. 16307, line 23:

> Q. Having had that opportunity to discuss this matter with Mr. Sunter and standing by your opinion, how does this opinion affect your opinion with regards to his approaches taken that we have seen summarized in HR-184 that deal with the gender bias issue?
>
> A. This was one piece of Mr. Sunter's evidence, this was one piece of his argument for not preferring the consultants to the

committees and there are other pieces to that argument. I don't have problems with the other parts that I have seen and I have indicated to -- I have given evidence on the other part, so I don't have problems with those pieces of evidence.

In general, despite the fact that I disagree with this part of his testimony, I don't have problems with the way he has handled the committees versus the consultants, even though I disagree with this particular step in his argument.

155

Q. I wasn't just referring to the committees versus the consultants, I was referring to the whole gender bias picture, all the other testing in HR-156.

A. I am restating that those analyses don't cause me a problem, no.

686. Shillington was asked to comment on Sunter's inclination to prefer committees to consultants, not for any statistical reason, but rather from Sunter's perspective that a decision of a group of individuals was preferable to a decision by an individual who may have had more advanced technical training. Shillington shared Sunter's opinion, and indicated he too preferred the consensus of seven people chosen in a balanced way, rather than one well-trained technical expert, at least on an issue like pay equity.

687. Both statisticians, Sunter and Shillington, agreed and informed the Tribunal that if we plan to use Sunter's t-test results to make adjustments to the evaluation scores of the committees, then the consultant scores are no longer simply a reference point but are, in effect, being preferred to the committee scores. In this context, the statisticians are of the opinion the consultant scores must be deemed to be free of gender bias and gender-related bias, before any adjustments are made to the committee scores.

688. Shillington testified the basis for his opinion was not statistical, but based on scientific reasoning and logic. His response is found in Volume 136, at p. 16706, line 14 to p. 16707, line 6:

A. I will leave it to you people to debate whether or not it's statistical. The question is if you are asking that you could use as a reference point for assessing gender-related preference someone who was consistent and unbiased, that wouldn't imply that

you are preferring those scores. I think that's the nub of the question here.

Q. That's right.

A. I am having problems with that because I just don't see the logic to it. I'm saying it's scientific reasoning. To me it's logic.

Q. Can I put it to you this way: Your concern is that you can't see how someone can apply a plan consistently without gender-related bias and yet not be preferred. Is that it?

A. Yes.

689. Willis defended the impartiality and objectivity of his consultants, and testified the consultant's re-evaluations can be used as a point of reference for determining a pattern of different treatment between male and female questionnaires by committees. He based his opinion on his belief his consultants had always followed a philosophy of not favouring

one side or the other, they had more experience in performing job evaluations and could evaluate consistently and without bias. Finally, the consultants had more experience interpreting difficult questionnaires.

690. Fred Owen, a pay equity expert, and a former consultant of Willis & Associates who participated in the JUMI Study, testified he believed it very important in determining the reliability of evaluations in the JUMI Study, that the consultants provide a frame of reference in order to determine the accuracy of evaluations. It was his opinion the consultant evaluations could be used as a standard for comparison for several reasons. His first reason is the consultants have an extensive knowledge and experience not only with the evaluation plan, but have a broad exposure to evaluations in a wide variety of jobs. His second reason is the consultants had access to an entire array of jobs that were being evaluated and that individual committees only had access to a smaller group. His third reason is the consultants had no knowledge of the Employer's classification system or pay ranges for any of the classes of jobs and did not have any preconceived ideas about the pay system. He testified the consultants themselves did frequent, almost daily, quality checks not only to determine how consistently the MEC discipline was being applied, but also to check the evaluations done by the consultants themselves to determine if the consultants were correct in their evaluation.

691. In Owen's written opinion (Exhibit R-167), confirmed by his oral evidence, he outlined criteria for adopting the committee evaluations. He suggested if the committees exhibit a good grasp of the evaluation plan as demonstrated by the reasonableness of their evaluations, and if there was no observable attempt on the part of any committee members to manipulate the evaluation outcomes nor to give prejudicial favour to any occupations or incumbents, there would be no need to assess committee evaluations against the consultant re-evaluations. In Owen's opinion, the evaluations fell short of these criteria owing to the lack of complete job information, as well as the observable behaviour on the part of some committee members who manipulated the evaluations so as to over-score female-dominated jobs and downgrade or under-score traditional male-dominated jobs.

692. There is ample evidence the JUMI Committee, during the operation of the study was prepared to use the consultants as a standard. The JUMI Committee had agreed to use the consultant scores as the baseline for comparison during the ICR testing. In that case, the consultants evaluated the test questionnaires that were provided to the committees, and the consultant scores functioned as a baseline for the ICR testing.

693. Throughout the study, the consultants were used by Willis as a standard to validate the committees' work. In a letter to Willis dated January 6, 1989, the JUMI Committee co-chairs requested Willis to provide baseline scores for the test questionnaires in the ICR and the letter reads in part:

> ...Your failure to provide baseline scores has delayed the work of the Inter-Committee Reliability (ICR) Sub-Committee as this information is necessary to analyze the consistency of ratings of committees with respect to a standard.

(Exhibit HR-82)

694. There were other occasions, during the JUMI Study, when both the management side and the union side jointly and separately requested the Willis consultants to review committee evaluations. Although this did not occur in the same framework as the ICR testing, in which the consultant scores were used as a baseline for comparison with committee scores, the consultants opinion, however, was sought as a check on the quality of the committee scores. Consultant reviews, with respect to the MEC benchmark evaluations have been previously described in this decision. There remains the agreement by the JUMI Committee to have Willis engage his consultant Wisner to do the 222 re-evaluations of the evaluation committees. There is

also the less formal reviews done by the consultants during the operation of the five and nine evaluation committees to test for consistency.

695. The following excerpts are further examples of consultant evaluations of committee questionnaires by Willis to validate the results, in Volume 60, at p. 7435, lines 3 to 23:

> Q. While the Master Evaluation Committee was performing their independent evaluations, were you also reviewing the questionnaires that they were looking at?
>
> A. Yes.
>
> Q. For what purpose?
>
> A. Part of the job is to, in effect, validate the consistency of their evaluations. My role, for the most part, would be to review the questionnaires along with the committee, to listen to their discussions and to do my own personal evaluation of the job based on the information that was brought forth. Then I would track that.
>
> While I did not give the committee my evaluation, I would track the consensus against my evaluation as a means of controlling and assuring myself that they were in fact being consistent in their interpretation of the information in the questionnaires and in the evaluation system itself.

Also in Volume 67, at p. 8429, lines 2 - 10:

> A. I responded to a number of concerns expressed and re-expressed by the Treasury Board from the summer of 1989 -- the summer of 1988 on. I had felt that we had put to rest the issue of whether or not the Master evaluation committee was evaluating fairly and equitably. I, in effect, validated the results. I said they were creditable and credible and yet the problems kept surfacing.

696. The JUMI Committee's reaction, during the study, to Willis' request to conduct the Wisner 222 did not, at that time, call into question Wisner's impartiality. It is reasonable to conclude the parties

themselves, at that time, assumed the consultants were bias free in performing their role in the process.

697. The parties understood from Willis there was no correct score to any one questionnaire. As the process continued, the only measure taken in the event of possible gender bias as contemplated by the parties and the consultant was to implement steps for improving the process. These steps or safeguards have been previously described. Having Willis counsel evaluators and provide additional training for either individual evaluators or committees occurred as part of these safeguards.

698. Willis testified the use of consultant re-evaluations after the process is concluded is quite different from their use while the process is ongoing. Willis testified after the process, the re-evaluations are used to identify whether or not there is a gender based pattern of difference. At the end of the study, Willis does not think it is particularly important to know the reasons for the disparities between consultants and committees because it is only the existence of a pattern that is important in his opinion.

699. Willis' firm belief is he and his consultants are without any kind of pattern in their evaluations. Willis states in Volume 210, at p. 27323, lines 9 to 12:

> A. It's my considered judgment that the experienced consultants with Willis & Associates tend to be bias-free or as nearly as it's humanly possible to be.

700. He went on to explain by "bias-free" he meant there was no differentiation on a gender basis between males and females. He was questioned as to whether he believed his consultants were without gender-related differences, such as hierarchical treatment where a consultant would be more liberal at the high end of a point scale or more conservative at the low end of a point scale. He responded as follows in Volume 210 at p. 27323, line 23 to p. 27325, line 22:

> A. That's an interesting point.
>
> Q. That one is a little harder to say, is it?
>
> A. Well, there is some evidence in a number of studies that we have done that it's difficult to get a good handle on a job that's two or three levels above your own. Alan Sunter made an observation that what might be viewed as gender bias might be something else.

Q. Yes, that's good. I'm going to talk to you lots about that point, so we don't have to -- bring me back to it later if I haven't dealt with it in detail. You say there are some studies to suggest it's difficult to get a handle on jobs three or four levels above your own, but your consultants were normally people who had very high-level jobs before they joined you, weren't they?

A. And they are consultants who have had some experience in evaluating higher level jobs. One of the problems in addition to it being difficult for a committee member to evaluate a job several levels above their own -- that is, having to have a good understanding of principles and theory and how is this important and what does strategic planning mean and things like this, things that are somewhat foreign to them -- and at the same time we find that the more complex jobs are more difficult to describe.

So, it's not unusual for -- I think it was Alan Sunter that suggested that perhaps the consultants had evaluated the higher level positions more liberally than the committees had.

Q. And that would be consistent. I gather what you are saying is that would be consistent with experience you have had in watching consultants and committees evaluate jobs?

A. I would say that would not necessarily be unusual.

Q. That's one take on it, that the consultants may be in a better position to appreciate those jobs. I would suggest the other factors at play with higher level jobs, I think I recall you telling us at one point that people tend to evaluate their own jobs more highly than they tend to evaluate other jobs that perhaps they are not as familiar with. Right?

A. I think maybe we are all a little bit biased in that direction.

701. Willis' rationale for not examining the reasons for the differences in the disparities between committees and consultants is because he believes it would be very difficult to pick out individual evaluations in order to explain the difference.

702. However, there were occasions during the study when Willis examined the consultants' (Drury and Wisner) evaluations to achieve an

understanding of the differences between the consultants and the MEC. Willis did this sort of analysis with 46 MEC benchmarks that showed differences between Wisner and the MEC of more than 10 per cent. In this analysis he was looking for a pattern. Willis' analysis involved an assessment of whether or not there was any pattern or apparent pattern of gender bias. He did this by reviewing the differences and the reasons for the differences as identified by the committees' rationales.

703. Willis agreed in cross-examination the difficulty when comparing differences between the consultants and the committees lies in determining how much of the difference is attributable to a particular factor, because there is no guarantee it is just one factor which accounts for the disparity. (Volume 210, p. 27350).

704. As to the differences in the way the committee and the consultant treated higher level jobs, Willis testified he was willing to accept the fact the consultants were probably more liberal in evaluating the higher

level positions. Based on his own experience, the consultants probably had a better understanding of the higher level jobs than the committees would. Willis gives a further opinion on this point in Volume 210, at p. 27355, line 18 to p. 27356, line 23:

> Q. You have told us why they might have a better understanding of them, but you will also agree with me that it is possible that in those situations where you have fewer benchmarks -- right?
>
> A. Yes.
>
> Q. And you have to exercise more judgment. Right?
>
> A. Yes.
>
> Q. -- that the consultant's view of those jobs might be influenced by their experience with high level jobs outside of the federal public service.
>
> A. And other studies which they have done. Yes, that's possible.
>
> Q. So you can see that there are things that might make them in a better position to have a "preferable" view of those high level jobs. Right?

A. I don't think there is any question about that.

Q. You have just told us that one thing could be that they could be influenced by things outside, by their baggage from outside studies.

A. I would say that when we are talking about high level, complex positions, the consultants should have a better grasp on the content of the job than any one of the evaluation committees that may not have had that kind of experience on their teams.

G. WHETHER THE RESULTS SHOULD BE ADJUSTED - THE EXPERTS

705. Willis testified the Tribunal has three alternatives in dealing with the reliability of the results: (i) to implement the study as it is; (ii) to adjust the results; or (iii) to trash the study.

706. As to option (i), Willis said without statistical analysis and the advice of a statistician he could not accept the results. In Volume 78, at p. 9576, line 19 to p. 9577, line 8, he said the following:

It is true that I was not happy with the various steps that were undertaken and to some extent we were able to do some shoring-up. However, without any analysis at all, without any opportunity to do some statistical analysis or to have it done and have some advice of a statistician, I don't think I could have accepted the results.

Once the study is complete, then it is possible to look at the results without regard to the other issues and make a separate determination: Do we have a consistent result or do we have a certain amount of bias and how much bias? In a sense, you do change into a different gear after the study is over.

707. With regard to the third option, Willis stated the following in Volume 78, at p. 9574, line 15 to p. 9575, line 7:

THE CHAIRPERSON: Could you tell us when the third option would be utilized?

THE WITNESS: I would want to sit down and talk to Milczarek and review all of the details with him. But it is possible, I assume, that the results would be so far out of line that they

just would not be believable. At that point, they should be trashed.

If we had stopped after the 222 evaluations, nothing had happened after that, and I were asked by the decision-makers what to do with it, given no opportunity to analyze the results, at that point I would say there is nothing we can do with it. We can't use what we have so far for any valid results. The 220 was too small a test by itself to make any judgments. So, if we aren't going to be able to do anything more, then we have to forget the study.

708. In his last appearance before the Tribunal in June of 1994, Willis testified, as he had done previously, he would rule out trashing the study. Willis suggested the study was about fairness in the treatment of employees, and the difference between the consultants and the committees, which resulted from Sunter's analysis was so small, in terms of a single employee's salary, "by the time you take out the income tax, that is not enough to pay for coffee." (Volume 211, p. 27489). On the other hand, Willis remarked "We are dealing with millions of dollars, so maybe there is more to it than just fairness to the employee." (Volume 211, p. 27489).

709. After having met with Sunter, Willis was interested in knowing how much of the difference between the committee and consultant was really a value bias. It was Willis' opinion if the value bias reduces the extent of difference between the consultants and the committees to the point where it is immaterial, no adjustment to the committee evaluations was necessary. Willis suggested if the difference between consultant re-evaluations and committee evaluations did not decrease after further analysis, then in view of the amount of money involved, "he would tend to adjust." (Volume 211, p. 27492).

710. Although no witness testified on behalf of the Employer concerning the Treasury Board's methodology paper (Exhibit HR-185), the evidence demonstrates that the Treasury Board made an adjustment to the evaluation scores by taking the Wisner 222 re-evaluations as a baseline. The adjustment preceded the Employer's equalization payments of January, 1990. The Employer adjusted all scores, other than the benchmark scores,

for which there was a consultant re-evaluation. The questionnaires were adjusted according to two regression equations contained in Exhibit HR-185, at p. 11, footnote 7. Shillington was asked his opinion on the regression

equations contained in Footnote 7 and responded in Volume 134, at p. 16401, lines 13 - 25:

> THE WITNESS: I would not adjust. I can tell you that when I first saw those equations and knew much less about the background to the data, I formed the opinion that I have expressed several times, that the onus is on the person -- before adjusting, I think there's an onus on the investigator to show the adjustment is warranted and the evidence here is that the adjustment does not warrant it and yet it was done. I formed that opinion as a statistician before I knew much more about the background of the study and nothing that I have heard in the background has changed that view.

711. Sunter expressed the same opinion as Shillington about the Treasury Board adjustments and said the following in Volume 106, at p. 12745, line 21 to p. 12747, line 10:

> The point about this is that the regression equations given in the Treasury Board document do not even approach the level of certainty that I would consider necessary to make any adjustments at all to the male and female evaluations.
>
> THE CHAIRPERSON: Could you explain that a little more.
>
> THE WITNESS: Because they are not significantly different. If I wanted to make an across-the-board adjustment on the basis of gender, I would have to be virtually certain of a number of things.
>
> One, I would have to be certain that the consultant is to be preferred to the committee, and I am by no means certain of that. As I tried to show yesterday, there are good reasons to doubt that.
>
> Second, I would have to be sure that the reason for the difference is gender, not something which is merely related to gender in some fashion.
>
> Finally, I would have to be sure of the numbers that I am using if I wanted to make an adjustment.
>
> We have seen that the order of magnitude of difference between the consultants and the committee, depending on which particular equation you use and which particular set of observations you use, is of the order of about 2 to 2.5 per cent. Nevertheless, we have

a methodology here that arrives at an adjustment of 7 per cent. How can that be?

The answer is that this regression analysis is a very poor, crude instrument for estimating the difference. Even if I were to believe all the other things, it remains a very poor instrument for making that adjustment because of the inherent uncertainty of the regression analyses themselves.

712. Durber, on behalf of the Commission, supported the results of the study without adjustment. His conclusion on the issue of reliability is contained in Volume 154, at p. 19167, lines 4 - 24:

A. My conclusion is that the parties were enormously successful in producing a body of excellent job information. They went to enormous cost and effort to produce evaluation results. They tested those evaluation results, we have seen, exhaustively, at least they were exhaustive and I am not sure of the results on us.

I am quite confident that the studies I have looked at fall short of the quality of work that we see in this particular study. I think the parties deserve a great deal of credit for what they have produced and certainly I had the confidence in those results to suggest that the Commissioners rely upon them in examining evidence of a wage gap.

I do not believe that there is what I would characterize as evidence of bias. My bottom line is that the results should be taken as they are and that any calculation of wage disparities ought to be based with a great deal of confidence on the job evaluation results.

713. Sunter has consistently maintained throughout his testimony no adjustment of the committee evaluation should be made. However, in response to questions raised by Willis and at the request of the Commission, he did suggest possible adjustment procedures to the evaluation results. We will elaborate more fully on these procedures in the event we conclude adjustment of evaluations is necessary.

VII. DECISION AND ANALYSIS

714. Throughout the JUMI Study, the Employer and the Alliance relied on the expert testimony of Willis to advance their positions. However,

during the hearing and in both written and oral argument, there was considerable debate between the Treasury Board and the other parties concerning this consultant's role in the re-evaluation of questionnaires and whether the consultant could be relied upon to produce gender bias free evaluations.

715. The Tribunal finds the position of the Commission and the Alliance particularly puzzling. Willis' impartiality was not an issue prior to this hearing. In its submission, however, the Alliance cited a number of reasons why the committee evaluations should be preferred to the consultant evaluations. It claimed, for example, consultant "baggage" and other factors such as age, sex, education, and lack of gender sensitization training which, it alleged, would contribute to consultant gender bias.

716. In our view, the Alliance was attempting to discredit the witness upon whose expert opinion it relied upon in terms of the data-gathering and job-evaluation process which occurred during the study. By way of further illustration, we refer to the following exchange between Counsel for the Alliance and the Tribunal which appears in Volume 224, at p. 29495, line 17 to p. 29500, line 19:

> THE CHAIRPERSON: Before you go on, Mr. Raven, I think I would like to respond to the word "antagonism" that you perceive from the Tribunal. I think it's a fairly -- it's a word that carries some connotation. I think that what the Tribunal has tried to do is understand your argument.
>
> These parties have put forward or engaged these consultants to assist them in conducting a study over a period of five years. When we are faced with an argument that these consultants could be gender biased -- I think that what the Tribunal has tried to do is understand and to challenge you on these types of arguments that you are putting forward. I don't think that our conduct in doing that -- I don't think it's fair to say that we're antagonizing or we're being antagonized, or whatever. I think that's our role and we will continue to do that role to try to understand and appreciate what it is you are trying to put forward to us.
>
> MR. RAVEN: I appreciate that. I was really attempting more to provide some added definition to my submission for that purpose.
>
> MEMBER FETTERLY: Before you start, I would like to make a couple of comments about this issue.

To begin with, if this were a civil trial and Mr. Willis was your witness -- technically he's not; he's the Commission's witness -- would you be permitted to discredit him after having introduced him as a witness?

MR. RAVEN: Mr. Fetterly, let me respond to the question in this way. I am not attempting to discredit Mr. Willis. That may be where we ---

MEMBER FETTERLY: You certainly give that impression, Mr. Raven. Let me just add this: Mr. Willis and his fellow consultants were the only experts who were actually involved in JUMI. You rely and he has defended the MEC results not only before this Tribunal, but also before JUMI. He has defended the ICR results. And he has done that both before this Tribunal and before JUMI. He has defended the total results before this Tribunal. Basically, he has said that they should not be trashed. And it's his plan that was adopted as being a gender-neutral plan.

To hear you and, to some extent, Ms. MacLean, attack, in a sense, his neutrality really puts the Tribunal in a very awkward position. I find it a matter of real concern. It's not a question of antagonism.

MR. RAVEN: What I had hoped to do this morning is try to clarify where we're going with this. Your comment, Mr. Fetterly, is very apt. It allows us and it affords me an opportunity to deal with that.

There is, in no sense, an attack here on Mr. Willis. That suggests that Mr. Willis or anyone associated with his firm was guilty of some malfeasance or misconduct in the way they conducted themselves in the course of the study or in the way they --

MEMBER FETTERLY: Not at all. Not at all. Mr. Willis and his associates hold themselves up to be experts in pay equity. They promote their plan as being gender-free. They train evaluators in order to evaluate on a gender-free basis. Now you are saying that their own ability to evaluate on a gender-free basis is suspect. That to me is a real contradiction.

MR. RAVEN: The fact that the Willis Plan was accepted by the parties here as being gender neutral for purposes of this study is

one thing. But if you will permit me to make this point, Mr. Fetterly, there's no personal attack on Mr. Willis or his associates. What we are trying to grapple with here is a very, very minute pattern difference between the consultants and the committees for the high end top quartile of male jobs, and we are now having to wrestle with the problem of whether we should adjust those scores to bring the committee scores in line with the consultants and whether there are compelling reasons to do that or not do that.

The submissions that are advanced here that I am about to get into is to raise with the Tribunal pertinent considerations in determining whether or not it makes a lot of sense to adjust in these circumstances. It's not intended as a personal attack on Mr. Willis.

MEMBER FETTERLY: That I understand. I think that's quite legitimate.

As I said to you yesterday, is it necessary, in order to achieve that, to allege that the consultants' ability to evaluate on a gender-free basis is suspect? Is it necessary for you to do that in order to establish or to argue that the committee results are to be preferred over the consultant results? I don't think it is.

MR. RAVEN: I tend to agree with you that there are a variety of reasons that support preferring the committees' scores, not just the questions that have been asked here and that are raised as to the manner in which the consultants themselves did these re-evaluations.

For example, if I understand Ms. MacLean's submission the other day, it was that the consultants had a slightly different discipline, a more liberal discipline, than the committees did.

Mr. Willis recognized that, and in his reports to the Joint Union/Management Committee, recognized that and found it quite suitable. In fact, he in his own words said "Given the context, our previous understanding and application of the Willis discipline in other contexts is not to be preferred to MEC's."

I don't know that that necessarily raises the question of bias, conscious or unconscious, or pattern differences. It does,

> however, confirm that (1) there were differences in the discipline that Mr. Willis has adopted in other studies and the MEC discipline; (2) that the MEC discipline was more conservative; and (3) the committee scores were more conservative than the consultants in high-end male jobs. So I don't raise that necessarily as an allegation of bias.

717. Statistical evidence was introduced by the Commission which, they submit shows inconsistency between the consultant Wisner, who conducted the 222 re-evaluations and the Gang of Four, who conducted the 300 re-evaluations. This evidence was also introduced in the context of whether committee evaluations should be preferred to consultant re-evaluations. In our view, it has a similar effect of discrediting the very expert the Commission contracted to do a further study and upon whom they relied during their investigation. Reference is made here to paras. 184 and 185 of Commission Counsel's written submissions:

> (184) A reasonable inference that the consultants as a group were not evaluating without gender bias or with relatively more gender bias than the committees may be drawn from the fact that Esther Brunet, a rater in the Willis II re-evaluations who was familiar with the federal public service, and considered a competent evaluator free of gender-bias by Mr. Willis, was almost 100% consistent with the committee evaluations (for the French-language questionnaires).
>
> (185) If an allegation of gender bias is supported by inconsistent application of the evaluation plan to male and female evaluations, then it is important to assess the relative consistency of the consultant evaluations compared to the committee evaluations. Consistency can be measured statistically. The statistical evidence of consistency of raters - committees versus consultants - demonstrates that it is the committees who are more consistent in their ratings than the consultants. The existence of a greater degree of rater error on the part of the consultants is described by Mr. Sunter as conclusive evidence that the committee is to be preferred over the consultant. Thus, the allegation of gender bias in the committee results is not supported by the statistics, nor is an allegation that the consultant scores are more consistent or more reliable.

718. We are of the view, there are other valid characteristics that can account for the differences between the Wisner 222 and the Willis 300 which should be considered quite apart from a pure statistical analyses. Although the two studies followed the same procedures, they are very

different in other respects. The Wisner 222 was undertaken to validate a process which brought Willis discomfort. It was a smaller study conducted by a single consultant who had demonstrated a more liberal discipline than the MEC. Wisner's analysis was a snapshot assessment only and was not intended to portray the whole picture. Not only was the time frame between the Wisner 222 and the Willis 300 different but the sample of jobs re-evaluated by Wisner 222 were from a smaller population than the Willis 300. The Wisner 222 were taken from the evaluations of the multiple evaluation committees and excluded the MEC evaluations. The multiple evaluation committees had been operating for about three months at the time of the Willis 222.

719. The Willis 300, was a larger scale study, undertaken after the process was finished. The purpose of this study was to confirm or to dispute the analysis contained in the Wisner 222. Four consultants conducted the Willis 300, with two or more consultants working in tandem. One of the consultants was an evaluation committee member. The sample of jobs came from the entire population of jobs from the expanded evaluation committees, excluding the Wisner 222. Not surprisingly, there was greater agreement with the committee evaluations in this latter study.

720. The timing of the re-evaluations by the so called "Gang of Four", the range of the sample, the number of consultants involved, the process followed and the circumstances then prevailing make the results, in our opinion, more likely representative of any real difference between the evaluation committees and the consultants.

721. The Tribunal had ample opportunity to observe Willis as he testified, during his first appearance which lasted 36 hearing days, and his second appearance which lasted 4 hearing days. We found Willis to be a credible witness who demonstrated patience, cooperation, and most importantly, impartiality in all respects. The Tribunal accepted Willis as an expert in the field of pay equity. Willis' experience, prior to the JUMI Study, was garnered entirely from his participation in U.S. studies in comparable worth. He had experience and had gained general recognition as a pioneer in this field. He was accepted as a qualified pay equity expert in the American court system.

722. We have reviewed the many occasions when the JUMI Committee asked Willis and his consultants to review committee evaluations or provide a baseline for comparison with committee evaluations. That role was well established and endorsed before the breakdown of the study. We do not now intend to view Willis' role differently from that which he provided to the

parties in the JUMI Study. All appropriate factors will be considered by the Tribunal if the issue of adjusting scores should arise.

723. The difficulties experienced by the multiple evaluation committees were not unexpected and should be accommodated and understood in the context of the sheer size of the Federal Public Service, its geographical dispersion and the multifaceted occupations and skills of its diversified workforce. These complicating factors coupled with the logistical problems which were encountered imposed a daunting challenge for all concerned. The experts, Armstrong and Durber emphasized the

difficulties inherent in the complex job evaluation process as it pertains to pay equity.

724. Given the nature of the JUMI process, the numerous participants with diverse backgrounds, and the working conditions within which the multiple committees functioned, the Commission and the Alliance submit job evaluation for purposes of pay equity will and must involve some conflict. This conflict, they submit, arises from a clash of values between evaluators who attempt, in a pay equity study, to question stereotypes and the attitudes of those with a more traditional mind set. Within that framework, the conflict which occurred is, (it is claimed), understandable and in fact unavoidable.

725. Respondent Counsel submits not all committees were working together in a "team effort" but instead operated in an adversarial mode. Willis said some committees "tended to feel themselves almost in a negotiation mode rather than a team of six or seven people trying to accomplish a common goal." Respondent Counsel submits it would be "wrong" for the Tribunal to accept the proposition pay equity job evaluation must inevitably involve conflict and adversity. Counsel submits pay equity job evaluation should be a cooperative problem-solving exercise in which evaluators work toward a common goal and evaluate based on the relevant facts. In the Employer's view, the process should instil confidence the relevant facts are being analyzed and that appropriate weight is being given to those facts. According to Counsel, when all these things happen, then the Tribunal can be confident the results are reliable.

726. In Weiner's opinion, the application of the plan is more important than the plan itself in ensuring gender bias free evaluations. She described the characteristics of the process which will prevent or minimize potential gender bias. In addition to having diverse committees of both genders and different organizational levels, Weiner stated other

factors, such as the training of committees, discussion as to how gender bias might operate, complete and up to date job evaluation information and the manner in which the committee conducts itself, must all be considered. On this point, she says in Volume 8, at p. 1092, line 13 to p. 1093, line 3:

> Q. Now, what about the way that the committee conducts its affairs on a day-to-day basis?
>
> A. Traditionally, job evaluation committees strive to be very efficient. They try to evaluate as many jobs as possible in a day.
>
> A pay equity committee has to take a different approach and open their questioning to asking for more information if they are unclear about something in the job information, to have a discussion about gender bias, to listen to themselves say things like, "This is just a secretary," and realize what they are doing, how this dismiss women's work.
>
> So all of those things take time, questioning, probing.

727. Weiner makes reference to "questioning, probing" in the context of committee evaluations. Although she did not comment directly on conflict in the committees, Weiner did insist that traditional values must be challenged in a pay equity job evaluation exercise.

728. The Tribunal is not persuaded, given the issue it has to decide that it should be asked to define the nature and degree of what is permissible, acceptable and legitimate discussion within the committee framework. Moreover, it is most difficult to measure its effect, especially when traditional values are being challenged and debated in a pay equity context. Nor is the Tribunal prepared to suggest answers for the resolution of conflict between committee members who may individually entertain strong opinions one way or the other on this sensitive subject. The study and implementations of equal pay for work of equal value in Canada is a relatively new discipline which is still in the developmental stage. Nonetheless, we do find it necessary, considering Willis' concern about committee conduct and individual evaluator behaviour, to assess whether the process achieved its purpose of producing gender bias free evaluations.

729. With regard to the effectiveness of the safeguards in place during the study, and more specifically procedures defined by Willis to be part of the Willis Process, we find the expert opinion of Willis to be most persuasive and informative. Because of its importance in assessing the results we have described in some detail the procedures and the safeguards which he recommended be adopted in that process.

730. The Tribunal believes it is incumbent upon it to comment on the JUMI Process as orchestrated by the JUMI Committee. Suffice to say, the JUMI Committee had a difficult working relationship from its inception. For incomprehensible reasons, the JUMI Committee chose to deprive both Willis and the Commission from real decision-making authority. This was done, notwithstanding the impartiality of both Willis and the Commission, their competence and broad experience in pay equity as compared with the parties themselves. In both the information gathering stage and in the evaluation stage of the JUMI Study, the JUMI Committee failed to follow Willis' advice and frequently refused to implement his recommendations. Some of the Willis recommendations were not implemented owing to "make or buy" decisions, largely controlled by the Employer and motivated by economic considerations. However, other Willis recommendations, not complicated by these considerations, were ignored as well.

731. Willis identified the JUMI Committee as a major weakness in the study and, in our view, his opinion is well-founded. The adversarial tone set by the JUMI Committee reflected the long-lasting and deep-rooted difficulties between management and union sides which permeated the JUMI Study throughout its entire life.

732. There is evidence the Chief of Pay Equity, an individual from the Treasury Board, viewed the JUMI Study in Willis' words, as a "bunch of bunk." (Volume 210, p. 27280). On the other hand, the Alliance wanted to follow a cohesive strategy as described in the correspondence from Millar, speaking for the Alliance, in announcing the Mont Ste. Marie meeting. This

incident and others threatened the foundation of the JUMI Study from the beginning and contributed in no small measure to the resulting difficulties. The union/management split was evident in the manner in which they attempted to resolve the issues. It even manifested itself in the seating arrangements at the JUMI Committee meetings with union and management on opposite sides. The parties opposed an attempt by Willis to change those seating arrangements. Willis said "...they looked at me like I was crazy." (Volume 60, p. 7459).

733. Willis disapproved of meetings the Alliance convened with their members prior and during the course of the study. There was the meeting of Alliance members at Mont Ste. Marie before the commencement of the study itself, where the subject of under-evaluation of female work was discussed in the absence of the consultants and the other parties to the study. During the course of the study, the Alliance also held evening meetings in which the participants discussed their logistical problems but during which there was also discussion relating to evaluations. Further, the Alliance representative on the MEC attended the evening meetings and was available to answer questions concerning the MEC benchmarks. At one week-end meeting, occurring in the fall of 1988, the Alliance held a training session on pay equity job evaluation, without the knowledge of Willis or the other parties. At that meeting, members examined and discussed certain of the MEC benchmarks. During this week-end meeting, gender-sensitization training, as interpreted by the Alliance, was given to the participants. The Alliance justified this unusual action on the grounds it was necessary to correct what it conceived to be historical injustices to women as "victims" in the work force.

734. Within the framework of the study, Willis felt he lacked the necessary support and backing of those in authority, both from the government and from the union sides while the study was ongoing. Although the sub-committee on communications had devised a strategic plan for communicating the JUMI Study to employees, Willis felt there was not enough emphasis on the need for communication from top management. He had initially proposed at least 10 consultant days for face to face meetings with department heads and union executives. No briefing sessions of this type were held and Willis believed this most likely resulted in the long delays before the employees completed their questionnaires.

735. Willis' evidence is that he designed the process to ensure a sound result, if the result is sound, it is immaterial whether the process is flawed. In examining the Willis Plan itself we find it to be an appropriate tool to evaluate jobs for the JUMI Study. During final argument, the Tribunal was informed there is no dispute between the parties concerning the Willis Plan. We refer to Respondent Counsel's written submission at para. 41:

> 41. Nevertheless, for purposes of this litigation, the Employer accepts that the Willis Plan was an appropriate plan to use in evaluating jobs in the Federal Public Service. Therefore, the Tribunal need not decide whether weighting of the Willis plan is valid.

736. We rely on Willis' expert opinion that the Willis Questionnaire, with slight modifications, was capable of capturing sufficient job information to ensure pay equity evaluation could be accomplished in the study. In his opinion the questionnaire contained sufficient information on which a well-trained and supervised job evaluation committee could provide reliable unbiased evaluations.

737. The degree of effectiveness of the safeguards provided for in the information gathering stage was disappointing to Willis. It was during this stage, that efforts to ensure the questionnaires were properly completed were made. Details of these efforts are described in the decision under the heading, The Willis Process.

738. In assessing the role of the coordinators we find, given the breadth of the study, it would have been extremely difficult for Willis & Associates themselves to act as coordinators without significant time delays and significant additional expense to the JUMI Committee. Coordinators were responsible for communicating directly to employees who were targeted to complete the questionnaires. Also, the coordinators trained incumbents as to the proper manner in which they were to complete their questionnaires. The consultants were involved with the JUMI Committee in the preparation of training materials supplied to and for the training of coordinators. If the number of completed questionnaires is a measure of the quality of the work of the coordinators, then their work can be viewed as most satisfactory. The percentage of return was impressive; nearly 100 per cent of the questionnaires were returned.

739. Willis' greatest concern lay in the lengthy delays in returning the questionnaires. According to Willis, delay in return of questionnaires impacts negatively on the quality of information, and the longer the delay the poorer the quality. There is little evidence as to what contributed or caused these delays. The evidence does not show the incumbents failed to fill out the questionnaires in a timely fashion and within the required 10 to 14 days after receiving training. Furthermore, there is little available information concerning when the coordinator-incumbent training sessions were held. To an extent, the large number of substitutions almost certainly contributed to the delays.

740. Although the effectiveness of the coordinators' role appears weak, this did not deter Willis from continuing with the evaluations. He was willing to have the study proceed notwithstanding somewhat weaker information. Willis instituted other safeguards, such as screeners/reviewers and the evaluators themselves to ensure completeness of job information. We do not consider the limitations of the coordinators' role to impinge significantly on the issue of reliability.

741. The screeners/reviewers applied a sophisticated technique of double check or safeguard. They were responsible for ensuring the questionnaires contained factually complete information for evaluation by the committees.

742. The screening and reviewing function was not conducted by Willis. Its sufficiency must be assessed from the training given, the evidence of

the witnesses who actually performed this function, the Commission's research (conducted by the outside researcher, Exhibit HR-245), and Willis' own observations and comments. The screeners/reviewers who testified believed they had done their job well. Through follow up telephone interviews they believed they were able to obtain the required information. Although Willis would have preferred more face to face interviews, overall he saw no difficulty with their performance or the role they played in the JUMI Study.

743. The screeners/reviewers received the same initial training on the Willis Plan as was given to the MEC evaluators. They also received "on the job training" from the consultant when needed. We find they functioned well and with no apparent problems other than the involvement of some committee "outliers" in this work. However, there is no evidence the "outliers", who bore this identification because they tended to evaluate differently than their committees, failed to perform their task fairly and competently or that they unduly influenced others. The six "outliers" who functioned as screeners/reviewers were relatively small in numbers compared to many others who fulfilled this role.

744. It is understandable why Willis would have personally preferred "hands on" involvement in the screener/reviewer function. However, it seems unlikely, given the volume of questionnaires, one consultant could have accomplished this task during the time frame allocated. Having carefully reviewed the evidence as it relates to the collection of job information, we accept Willis' opinion and find as a fact the job information was of satisfactory quality when all the "shoring up" is taken into account.

745. Consistency is an important feature in the process of pay equity job evaluation. The Willis Plan should be applied consistently especially when multiple evaluation committees are involved. This requirement, if met by the participants, does not necessarily imply the process is without gender bias and, on the other hand, lack of overall consistency between the committees does not necessarily imply that the evaluations are biased, nor

is it crucial to the issue of reliability. In the final analysis, Willis' concern was whether the results were biased. However, within the context of this study and in assessing how well the process worked, we consider it prudent to comment on whether the multiple evaluation committees consistently applied the discipline established by the MEC.

746. There were some committees amongst the original five evaluation committees, namely Committees #1 and #2 and the first version of Committee #4, that worked well. After the restructuring of the original five multiple committees into nine multiple committees, the newly created nine committees appeared on the whole to have functioned well. Most of the multiple evaluation committees did, in fact, attempt to follow the MEC benchmarks, adhere to the discipline created by the MEC and follow the same job evaluation procedure as had the MEC. There is evidence, at least from the early ICR testing, of consistency between committees in interpreting the Willis factors and applying the plan. To some degree, the MEC benchmarks had a steadying effect on the functioning of the multiple evaluation committees and on the study as a whole. This is most evident

from Willis' response to a question by the Tribunal regarding the first incarnation of Committee #3 in Volume 69, at p. 8676, lines 8 - 18:

> But, as it worked out, one of the things maybe that helped to stabilized [sic] the evaluations was that we did have those Master Evaluation Committee benchmarks for them and maybe they just got so tired each fighting for their own side that they went along with the Master Evaluation Committee's benchmarks. I was not at all satisfied that I could leave it at that or let it rest at that. But I could not observe any particular problem in the actual evaluations that we were able to examine.

747. The Tribunal will now refer to the training the committees received in order to properly perform their function as evaluators. Willis' approach in dealing with gender stereotypes and traditional values is to direct evaluators to break down a job into its component parts and to evaluate each part separately so as to ensure bias free evaluations. Willis' opinion differs from Armstrong's about whether his method of training should have included a more formal kind of gender sensitivity training which would focus on under-valuation of female work. In our view, the fact this training was not formalized by Willis does not increase the potential for gender biased evaluation. Willis preferred "on the job training" and this approach was used by him successfully in previous studies. Moreover, the JUMI Committee had authority to decide what was to

be included in the training and what training it expected to be provided. Willis was criticized by the Alliance, during this hearing, for not providing gender sensitivity training in the form espoused by Armstrong and in the reference material from the Ontario Pay Equity Commission. It should be noted however, that the Alliance approved of Willis' training approach at the outset of the JUMI Study while it was a member of the JUMI Committee. The Alliance's criticism of Willis would seem to be motivated by Willis' disapproval of the Alliance undertaking this kind of training during one of their meetings held in the absence of the consultants and the other participants. In addition, Willis commented on another aspect having to do with the quality of such training in Volume 211, at p. 27483, line 24 to p. 27484, line 20:

> Q. On another subject -- and this is one that you have discussed at some length with my friend Mr. Raven. It's the subject of training participants in a study to be sensitive to gender issues. Do you recall the subject?
>
> A. Yes, I do.
>
> Q. In deciding whether such training is beneficial, is it relevant to know something about the quality of the training?
>
> A. Certainly.
>
> Q. Could you comment on that, please?
>
> 174
>
> A. I would think it would be important for whoever is providing the training of this nature to be accepted as an impartial individual and to have been trained in this area.
>
> Q. If the training is not done well or impartially, could it have any effect other than off-loading baggage?
>
> A. It's possible that it could have the effect of creating more baggage.

748. We hold the view, in recognizing Willis' extensive hands on experience in conducting pay equity studies, that his practical approach has merit and is acceptable. We say this notwithstanding Armstrong's opinion based, it would seem, entirely on research and on the available literature.

749. With respect to the actual job evaluation process, there is anecdotal evidence the process did not work as well as it ought to have. Willis testified about his discomfort with the behaviour of some committee members, particularly with the first version of Committee #3, which he characterized as consisting of "two warring camps". He was thwarted by the JUMI Committee from taking the appropriate remedial measures he believed were necessary concerning those evaluators who were evidencing gender bias on Committee #3.

750. The Tribunal had the benefit of observing and hearing witnesses who had participated in the evaluation committees. Their evidence can be characterized generally as an 'injection of reality' into the evaluation process which is best described as a lengthy, arduous, complicated, stressful and difficult process. In general, these evaluators did not express difficulty with the sufficiency of the information provided in the questionnaires. If and when further information was required by a committee to complete an evaluation, this was accomplished through the procedural safeguard established for that purpose, that is, having the screener/reviewer supplement, clarify or obtain new information.

751. Willis testified about some of the strengths of the JUMI Study. He regarded three strengths of the JUMI Study as being, firstly, the large number of individuals who participated on evaluation committees, secondly, the large number of diversified jobs evaluated and thirdly, the large number of jobs in the sample which enabled him to deal with "slightly greater disparity" in job information than a study with a smaller population. Willis believed the committees represented a "pretty" good balance of union and management employees with different backgrounds despite the difficulties the unions encountered in naming male representatives. There is evidence some of the female evaluators were members of male-dominated unions which contributed to more diversification within the committees.

752. One of the problems Willis recognized was the participation of management individuals trained in classification. Seven evaluators nominated by the management side had extensive knowledge of the classification system in the federal government. They served on four of

the evaluation committees and on the MEC. The problems associated with classification backgrounds surfaced during the evaluation process. The statistical evidence, however, did not identify the classification background of these individuals had an impact on the multiple evaluation committees' consensus scores. There is anecdotal evidence these

individuals had little or no influence and tended to be ignored by the other participants.

753. Another problem which arose was the participation of some Alliance supporters who evidenced an agenda for increasing the value of female-dominated jobs. There were misguided attempts to influence the evaluations of some of its members through confrontation and intimidation. The quantitative differences in the consultant re-evaluations point to the committees under-evaluating some male-dominated jobs but do not demonstrate these misguided individuals accomplished their objective of persuading others to over-evaluate female-dominated jobs. As Sunter's analysis reveals, significant differences between the committees and the consultants exist almost entirely in the treatment of male-dominated questionnaires. Furthermore, the IRR test results reveal the majority of both management and union outliers exhibited a male preference. Thus, any conscious attempt by Alliance members to over-evaluate female-dominated jobs was unsuccessful. There is also some comfort to be had in the testimony of all of the Alliance evaluators who gave evidence to the effect there was no Alliance meeting at which members were told to over-evaluate female-dominated jobs or to under-evaluate male-dominated jobs.

754. Some of the evaluators were identified by both the consultants and the IRR test results as outliers. During the JUMI Study, efforts were made to assess whether the outliers were exercising influence on the committee's final consensus. The statistical analysis demonstrated their influence was negligible. As well, while directly observing the participation of the outliers in the evaluation committees, Willis could not detect them exerting any influence on the other members.

755. One of the most redeeming features of the JUMI Study was the work of the MEC which had the unqualified endorsement and support of Willis. When the MEC completed their work, Willis was satisfied they had done a good job. There were several reviews of MEC's work by the consultant, revealing some differences between the MEC and the consultant evaluations. Willis was not concerned with the extent of these differences, as there was no evidence of gender bias in the MEC evaluations. Willis said he anticipates differences between committees and consultants. In his view, the presence of those disparities does not necessarily mean the consultant is "always right".

756. From Willis' perspective, there are four questions that need to be addressed in deciding whether or not a real problem exists. They are:

(i) What is the extent of the disparities on total scores in a specific evaluation;

(ii) How frequently do the disparities occur;

(iii)The rationale: why have the committees done what they have done; and

(iv) Is there a pattern to the disparities, and if so what is the pattern?

757. When the study is over, Willis examines the total score, to answer two of the above four questions, namely, what is the extent of the disparities and how frequently do they occur. Willis' examination is done with the assistance of a statistician, upon whom he also relies for the answer to the fourth question, namely, whether there is a pattern in the disparities. There can be a number of reasons for the disparities referred to in question (iv) but, at this stage of the study, Willis is not interested in those reasons.

758. In Willis' opinion, when the study is completed, the appropriate consideration is how much did the committees stray from the consultant evaluations. In his view other considerations are, at this point, immaterial. His reason for considering only the bottom line results is that the evaluation committees are no longer functioning. An understanding of whether or not the committees were applying the plan correctly is no longer useful to the consultant because counselling and training is no longer feasible.

759. Willis expressed the view, on a number of occasions during his testimony, that the results were more important than the process. By results, he meant the comparisons between the committee evaluations and the consultant re-evaluations.

760. However, in view of our interpretation of s. 11 of the Act, which is that causation is implicit in the legislation, we must address the question of whether the differences between the consultants and committees arising during the process are based on gender, or on some other consideration. It follows therefore, it is not only necessary but crucial that the evidence be examined in detail in order to determine whether or not the differences between the committees and the consultants are gender based.

761. There was evidence led by the Alliance concerning analyses done by two individual Alliance witnesses who examined committee and consultant rationales, with a view to explaining consultant and committee disparities.

Prior to the commencement of the evidence of the first of these witnesses, the Employer provided an admission to the Tribunal which reads in part:

> 4. The Employer makes the following admission and clarification in order to narrow the issues and to avoid further unnecessary use of hearing time in tendering evidence.
>
> 5. The Employer admits that disparities between consultants and committees in the Wisner 222 and Willis 300 re-evaluations may have occurred for reasons other than gender bias in the Joint Initiative Committees.

> 6. To clarify the issues, the Employer will not rely on the reasons for disparities as evidence of gender bias in the process or bias in the results.
>
> 7. Therefore, the Employer contends that evidence analyzing the reasons for disparities does not assist the Tribunal to assess:
>
> (a) the reliability of the process; or
> (b) the reliability of the results.
>
> (Exhibit R-154)

762. Willis had an opportunity to comment on the two analyses presented by the Alliance witnesses. Willis does not consider either of them helpful for identifying gender bias in a large study or for exploring consultant and committee disparities. In his experience, individual assessments of differences based on the rationales will not reveal the existence of gender bias. The Tribunal accepts Willis' view. Our determination will not be based on what is contained in the rationales for individual differences between committee evaluators and consultants on a given question, but instead will be based on an examination of all the evidence relevant to committee and consultant evaluations.

763. Willis wanted questionnaires that were complete and focused on factual information. Incomplete questionnaires lead evaluators to make assumptions which result in a wider range of possible disparities. The number of disparities in this study tended to be higher than what Willis usually experiences. On the other hand, Willis had never before participated in a study as large as the JUMI Study and was not in a position to supervise the entire 522 re-evaluations, some of which had been done during and some after the study was over.

764. We will now address Willis' questions (i), (iii) and (iv). Willis testified on numerous occasions about a tolerance level of differences between committee and consultant evaluations. The percentage variances he uses are simply a function of his experience and what he views as acceptable. Based on the quality of information available to the MEC, he would expect to find a 10 to 12 per cent random variance, either positive or negative, in evaluations. Because the information available to the multiple committees was not, in his opinion, of as high quality as was available to the MEC, he would expect to see between 15 and 20 per cent random variance in their case. There is more opportunity for evaluators to make assumptions when they are furnished with poorer quality information.

765. Willis testified random variance occurs when value judgments are made about the meaning of the facts presented in the questionnaire. Willis considers in a large study, such as the JUMI Study with the sheer numbers of jobs being evaluated, greater disparity is acceptable as a result of the relatively weak job information. Willis is concerned, if over time, the variance is no longer random and becomes systematic. He defines systematic variance as value or values which are "consistently higher or lower than an objective evaluation of certain types of jobs." He treats the term "systematic variance" as equivalent to "gender bias".

766. Shillington testified on the distinction between pattern and randomness in a large study and the difficulty in defining something as random. He said in Volume 86, at p. 10540, line 9 to p. 10541, line 13:

> Q. How do you know that you have something that is random as opposed to something that is not, something that is patterned?
>
> A. Sometimes you are comfortable using a term without trying to define it, and randomness is one of those terms that is easier for people to use comfortably. I think everybody knows what you mean, but as soon as you try to define it, it gets difficult.
>
> If you show someone a pattern of numbers, quite often people will look at that patten and you can say, "Is it random or not?" It is very difficult to show that a pattern is random. It is often easier to show that it is not.
>
> Let me write down a sequence. Suppose we toss a coin four (4) times and we get heads, tails, heads, tails. You can look at that and say that that is a possible outcome from a fair coin. You have fifty (50) per cent heads and fifty (50) per cent tails. But

if you continued getting heads, tails, heads, tails, heads, tails, heads tails, something in our brain starts saying that this isn't random any more. Yes, you are getting half heads and half tails, but that is far too systematic.

Defining what is random is very, very difficult. It is much easier to say, 'This is not random. It looks like there is a pattern here."

767. He further states in Volume 86, at p. 10543, lines 1 - 8:

So, it is easy to show that it is not random, that there is a sequence. But proving it is random is virtually impossible.

We use the term "random" basically as a catch-all phrase for what we don't know. If you toss a coin over and over again, we say that the coin is random because we can't predict well the next outcome.

768. Willis confirmed at the conclusion of the study, he is willing to accept a wide disparity in evaluations provided there is no pattern. He does not like to see any pattern at all. He said in his earlier testimony if the variance is less than 2 per cent, he probably would not adjust the evaluations. He said in Volume 61, at p. 7596, lines 5 to 11:

A. In the final analysis when the study is over, obviously in many cases we are involved in recommending and implementation. At that point I might decide that there needs to be some adjustment to correct. But obviously, if it is less than 2 per cent, the difference in pay is so minimal that I guess I would have to accept it.

769. As a rule of thumb, even with the very best job information available, Willis expects to see more than plus or minus 10 per cent disparity between the committees and the consultants. Willis considers disparities over 10 per cent a "red flag" which suggests there may or may not be a real problem in the evaluations. In a large study, such as the JUMI, Willis seeks the assistance of a statistician to determine whether the disparities are systematic.

770. The nature of this exercise, which Willis describes as more an art than a science, renders it difficult to quantify job evaluation either statistically or mathematically. The Tribunal was occupied for a

considerable time with the presentation of statistical evidence. In the end, we had opinions from the statistical experts, Shillington and Sunter, to the effect that statistical analysis cannot identify the existence of gender bias.

771. Sunter's conclusions are a product of hypothesis testing. In his interpretative analyses, he relies on probability criteria and mathematical models to explain variations in the data. His conclusions are not based entirely on scientific reasoning and mathematical applications but, in part, on assumptions about the "nature of the world". Sunter repeated at different times in his testimony when his intuition assisted him in reaching his conclusions. The following examples, which are not exhaustive, are reproduced. In Volume 110, at p. 13221, lines 8 - 17, he remarked:

> When I said that the original stuff is most unexpected it was because I felt that if the consultant is always right and the committee is always wrong, then my statistician's intuition tells me this should lead to a larger variance for committee scores and it should lead to a negative covariance and a negative correlation between difference and committee scores, which is exactly the relationship that you see reproduced by Model 2.

772. As well, in Volume 119, at p. 14387, lines 10 - 20, Sunter said:

> There is a stronger, positive association between DIFF and CONS than there is between DIFF and COMM. Now, let me say that my statistician's intuition tells me -- I don't have to justify this, it's just that one develops an intuition, and my statistician's intuition is surprised by this, if it really is the consultant who is in error -- sorry, if it is the committee who is in error. I would expect the associations to be somewhat different, but I am just speaking intuitively now.

773. Also in Volume 123, at p. 15046, line 19 to p. 15047, line 2, Sunter said:

> I think he asked whether they were relevant tools in the context of what Dr. Shillington was doing in the IRR, and I said "yes". You know, he was in a different situation, concerned with different things, and I would assume that he used both of those

tests as a result of some kind of intuitive assessment -- which, under the circumstances, he was perfectly entitled to make...

774. And once again in Volume 217, at p. 28225, lines 9 - 23, Sunter remarked:

Typically, in decisions theory, with decisions, you associate losses and gains with various decisions, and how you make a decision is a consideration -- if you wanted to do it technically, you would have to go into all that stuff, and I am trying to skirt over it and say, "I have no loss function to offer here. I don't know how you should make that decision." If you challenged me to come up with one, I suppose I could, a decision-making function here.

This is why I am not taking a position on it. Make the adjustment or don't make the adjustment -- it depends on your kind of intuitive decision-making process, but I am not about to make that decision for you.

775. Both statisticians agree statistical analysis can lend weight to the evidence even though it may not be conclusive in itself. Shillington discusses significant and non-significant results in terms of weak or strong evidence. In his opinion, a significant result is not conclusive in itself. It may, however, lead a statistician to conclude a hypothesis is suspect or the statistician may draw an inference which casts doubt on the hypothesis. In Sunter's opinion, statistical analysis will lend weight to something which already seems plausible. The analysis can very seldom by itself provide plausible explanations. In fulfilling this limited role, we believe statistical analyses are appropriate and helpful. Therefore, we conclude, statistics are ancillary to the primary function of the evaluators to render a value judgment, and of the Tribunal, which is to determine the reliability of the results.

776. In Sunter's last appearance before the Tribunal, he agreed there were limitations to the applicability of statistics for the determination of the issues before the Tribunal. This is found in Volume 217, at p. 28301, lines 13 - 22:

MEMBER FETTERLY: I guess the point that I am trying to get at is this: Statistics don't necessarily tell us the whole story. I think you might agree with that, would you not?

THE WITNESS: Yes, I would agree with that as a general observation.

MEMBER FETTERLY: So we may have to consider other factors that perhaps are not within the realm of your speciality.

777. Sunter's tests help identify the statistically significant differences between the Wisner 222, the Willis 300 and the combined database (522) compared with the committee evaluations. Sunter interprets the differences as not having a consistent pattern. He found significant

differences between the consultants and the committees in both studies in the male-dominated questionnaires, but more so in the Wisner 222 than in the Willis 300. The results of his tests identified differences found mainly at the higher end male-dominated and some few higher end female-dominated positions. Overall, the female-dominated questionnaires had a lower distribution in value than the male-dominated questionnaires. We are mindful of the fact the differences with the female-dominated questionnaires were not statistically significant.

778. Shillington provided an opinion regarding Sunter's analyses of "other possible causes" for the differences between the consultant and the committee scores, that is to say, other than gender differences. One of these analyses included comparisons to determine if the differences were associated with the relative distribution of questionnaires in the higher and lower point ranges. Contrary to Sunter's view, Shillington was of the opinion it would be very difficult to separate out these two data analyses questions as to whether there is some reason other than gender which is the cause of those differences. On this point, Shillington says in Volume 131, at p. 16045, line 21 to p. 16046, line 21:

A. Yes, and the analysis that is behind that.

The regressions were done in a way to try to see if there was a relationship between the differences between the consultants and the committee in gender. It is also possible that any differences that might have existed between the consultant and the committee scores were not directly related to gender but perhaps were related to high values versus low values. This has been talked about here.

The confounding is introduced because there is a strong trend in the data for the male questionnaires to all have high values relative to the female and the female questionnaires have a fair tendency to come from the lower end of the spectrum, which means

you cannot separate those two data analysis questions, or it is difficult to separate them.

THE CHAIRPERSON: What do you mean?

THE WITNESS: You can't separate the question whether or not a pattern is related to gender or whether or not it is related to whether or not the scores were high or low.

779. On the same topic, he says in the same volume at p. 16048, line 16 to p. 16049, line 11:

In this circumstance, back to the analysis of the Willis scores and the possible adjustment, we have a situation which -- to the extent that there is a pattern here, if someone came and said this is possibly not due to gender, maleness or femaleness, but rather could be due to professionalization or some questionnaires having much higher values than others, you would have a problem extracting those two separate hypotheses from the analysis because

you have a situation in which the males predominantly had high values, the females predominantly had low values. So maleness is confounded with high and low values.

That is reflected in the distribution. That is why it is a distribution question. The distribution of the Willis scores for the males tended to be quite a bit higher than the distribution of the Willis scores for the females. It is a confounding issue. That is why in interpreting it you are going to have to be cautious about that.

780. In the end, Shillington suggests these analyses should be used with caution, and we refer to his response in Volume 131, at p. 16049, line 20 to p. 16052, line 7:

THE WITNESS: It is more of an interpretation issue and, I think can't be stronger than -- I am not Mr. Sunter, but I think that we have to make sure that when we use these analyses, because of the differences in the distribution, we have to be cautious.

THE CHAIRPERSON: For example, when we compare regression lines, we usually look at the differences -- or we have been looking at the wage gap using regression lines, for example, in calculating a

distance between them. So you are comparing them to see what is the distance.

THE WITNESS: Yes.

THE CHAIRPERSON: That is what I think when somebody says to me that you can't compare these two regression lines. So when Mr. Sunter is saying that you can't compare these two regression lines, I am saying compare them for what? That is why I am a bit confused.

Are you saying you can't interpret them, meaning that because in the male regression line you have distributions of both, high and low distributions, but a tendency to be higher, whereas in the females you have a distribution of a low and high but a tendency to be lower, but when you interpret these lines you can't say it is definitely associated with a gender-related bias, for example?

Is that what you mean?

THE WITNESS: Yes. I think it is more of an interpretation of whether or not the patterns that you are seeing are clearly related to gender or whether or not those patterns are related to high score versus low score because they are, in the data, occurring together. The males are predominantly high score and the females are predominantly low score.

THE CHAIRPERSON: So it is not comparing them in terms of calculating a wage gap. Is it?

THE WITNESS: I think that is a different issue which we will get to, I think.

THE CHAIRPERSON: Okay. But just looking at these and what you can say about what they describe in terms of their distribution, what you can interpret from that is that the males tend to be high, the females tend to be low, but you can't, because of this confounding effect, you can't really interpret anything else with certainty. Is that ---

THE WITNESS: That is right. You have to be very careful when interpreting the results because you have to keep in mind that if somebody came with an alternative explanation for the data and the

explanation was that this had nothing to do with gender, that this was high score/low score effects, you have collected your data in such a way that most of the high scores are males and most of the low scores are females. So they are two equally valid explanations for the same data.

I think it is a caution in interpretation that I think is reasonable.

781. Sunter conducted further analysis for presentation in reply. He refers to this analysis as his "value effect" analysis which attempts to explain further the difference in treatment of high point value and low point value questionnaires. The two statisticians hold opposing views as to whether such questions as value effect and gender can be separated out or "unconfounded". We note Shillington's warning to exercise caution when attempting to unconfound the data in these circumstances. However, the analysis is useful in demonstrating the differences between the consultants and the committees occur at the high end of the point range. Having found the applicability of statistics for the determination of the issue before us to be supportive rather than definitive, we are not convinced as to the necessity for, or the validity of, Sunter's other conclusions pertaining to his "value effect" analysis. Moreover, Sunter's earlier work which focused on identifying significant differences remains helpful and useful in understanding where the differences occur between the committees and the consultants.

782. We will now address Willis' question (iv). The Wisner 222 was completed while the study was still ongoing. At that time, Willis did not perform any in depth analysis to determine the reasons for the differences between the Wisner 222 re-evaluations and the committees' evaluations as he had done with the previous consultant re-evaluations of the MEC benchmarks. Willis would have preferred to proceed immediately with the second part of his plan, which was to do a larger study. He believed this further study was desirable because the Wisner 222 was inconclusive on the question of gender bias. It was recognized at the time by Wisner himself that there could be other plausible explanations for what he defined as an observed pattern of evaluation differences in the Wisner 222 (Exhibit PSAC-4). Wisner does in fact suggest positions in male-dominated classifications, with more complex duties and responsibilities, might have been the cause. He states:

...Because this is true, the observed pattern of evaluation differences could occur if the committees tended to under evaluate

more complex positions, in relation to the MEC discipline as seen by the consultant.

(Exhibit PSAC-4, p. 8)

783. It should be noted during the MEC's work Willis observed that the MEC adopted what he described as a "conservative" discipline. This is evidenced by the reluctance of the MEC to evaluate jobs above a certain level. The Willis evaluation plan had a varied level of complexity from levels A through to level G in functional job knowledge. According to Willis, the high G level is a level that presupposes "...a requirement for an expertise or command of a professional sphere of knowledge." (Volume 35, p. 4448).

784. During the operation of the MEC, Willis felt there were four or five questionnaires which should have been evaluated at the G level. Willis tried, at a special session with the MEC evaluators, to encourage them to promote jobs beyond the F level. Willis testified in Volume 35, at p. 4448, line 19 to p. 4450, line 24, about the phenomenon he observed:

A. Out of the ones that the Master Evaluation Committee had evaluated. As we got toward the end, in fact, I even had a special session with them to see if we could break out of the high F into the G level. It was an interesting phenomenon. They all realized the problem, but they just could not seem to select any jobs to promote above that F level.

In fact, I said "let's just pick one" -- I want the other committees to feel that they have a highly professional job with true expertise. I don't want them to feel they can't go beyond the F level. "So, pick the strongest job you can. Let's see if we can't promote it to the G level." And they just couldn't do it.

This was of some concern to me. That was mitigated, however, for two reasons: (1) there were several jobs at the high F level. The point totals for the high F are the same as for the light ---

THE CHAIRPERSON: Excuse me, you were saying there were several jobs at the high F level?

THE WITNESS: Yes, the F leaning toward G.

If you recall from the evaluation system, the G on the light side leaning toward F has the same point total. So I was not concerned from the standpoint of the points. But since they were the

committee that was setting the frame of reference for the other committees, I wanted them to be able to exercise that G level. That didn't happen with the Master Committee.

185

As it worked out, I was in counselling later with the evaluation committees. I explained the problem to them. I don't remember how many jobs they ultimately evaluated at the G level, but I understand that they did break through and they did evaluate some of the 4,000 at the G level.

Q. So was this tendency in the end something that you felt was beyond a concern?

A. The other mitigating factor was that even though they were very conservative here, this conservatism was consistent. Looking at the alignment, I felt that the internal alignment was still appropriate. So while they were very conservative at the top, it did not create, let's say, an inversion in the evaluation relationships.

There were so few jobs -- and I remember discussing this with Paul Durber after the study. They looked at those jobs that might have gone to a higher level and there were so few of them that they wouldn't have affected the results materially.

785. Early in the process, specifically with consultant re-evaluations of the MEC positions, there is evidence the consultants were evaluating differently than the committees. This first occurred when the Willis consultant, Drury, did her review of the MEC's evaluations at its own request. It also occurred later on, during Wisner's review of challenges to the MEC evaluations. Wisner's discipline was noted to be slightly more liberal than the MEC. Willis testified to this effect in Volume 56, at p. 6940, lines 14 - 24:

Q. So, this goes back to your comment that Mr. Wisner was probably more liberal.

A. He was slightly more liberal, but that didn't bother me. I had a reason for not wanting to do the evaluations myself or to have Jan Drury do them, even though we had discussed the Committee, I was willing to accept the fact that Mr. Wisner's discipline might be slightly different. But it was the

consistency in evaluation differences that I was looking for. So, Wisner made the best choice.

786. Willis was willing to acknowledge Wisner's discipline might be slightly different from the committees'. This did not concern him as long as there was no pattern in the differences.

787. Some evaluations were easier to do than others depending on the information in the questionnaire. Willis testified the responses from incumbents in female-dominated occupations were returned more quickly and contained better information than from incumbents in male-dominated occupations. He was asked if this could have an effect on the reliability of the evaluations in a restricted sense. His response is contained in Volume 68, at p. 8575, lines 3 - 13:

Q. But what I am trying to get at here is: Could that affect the reliability of the evaluations based on, let's say, occupational groups? In other words, were you getting more reliable information from predominantly female groups and less reliable from predominantly male groups.

A. I haven't tested that, but I believe that is a possibility, certainly, since the quality of the information does generally tend to be better from female-dominated groups.

788. Willis further testified the questionnaires from incumbents in high level technical and professional jobs were slower in returning and they contained weaker information than questionnaires from incumbents in clerical and vocational jobs. In this regard, Willis explains "generally speaking" the professional and technical jobs are more difficult for the evaluators to understand. He says in Volume 69, at p. 8582, lines 11 - 20:

THE WITNESS: Professional and technical level questionnaires would be less easy to understand than, say, trades or clerical.

MR. FRIESEN:

Q. And that is partly because they were not as well described in the information.

A. Partly, and partly because it is more difficult to understand a more complex job. [emphasis added]

789. Willis' opinion is verified by the testimony of at least two evaluators. Crich, a member of the first version of Committee #5, testified her committee had difficulty evaluating questionnaires from male-dominated occupational groups. In her view, this contributed to the problems experienced by that committee. We also have testimony from Latour, also a member of Committee #5, as to the difficulty this committee experienced in evaluating technical jobs.

790. For the most part, the QA Committee's work must be discounted in view of Willis' criticisms of their work. However, the evidence of two of the participants, Crich and Yates, merits consideration because it illustrates the difficulty experienced by the QA Committee members when evaluating the 25 male questionnaires identified from the Wisner 222 and their inability to achieve consensus in those cases.

791. By way of contrast, the consultants did not experience the same difficulty in evaluating the more complex questionnaires as did the committee members. The consultants had the benefit of professional job evaluation experience and training, which enabled them to evaluate those positions more easily than the committee evaluators. The fact the committee evaluators lacked that kind of professional expertise contributed, we believe, to the inefficiency of the job evaluation process and the lengthy discussions which took place during the evaluations.

792. Willis expressed a high regard for the competency and experience of his consultants in conducting pay equity job evaluations. Willis agreed

his consultants were more liberal in evaluating higher level positions. Considering the consultants' experience, background and education, he also believed they probably had a better understanding of the higher level jobs than the committee members.

793. Illustrations provided by the Employer were confirmed in the cross-examination by Respondent Counsel of Sunter and Shillington as to the effect of different treatment of female and male questionnaires on the wage gap. Different treatment (arising from gender bias) will have a direct impact on the wage gap. There are two distinct ways in which the wage gap will increase. An increase can occur if committees are under-evaluating male-dominated questionnaires. It can also occur when the committees are over-evaluating female-dominated questionnaires. In either case, it will have the same effect. Expressed in another way, the wage gap will be "over-stated" when either of these events occur.

794. If the 2.3 per cent disparity between the committees and the consultants is attributable to gender bias, then it arises either because the evaluators were consciously or unconsciously treating male-dominated jobs less favourably than the consultants or, on the other hand, were over-valuing female-dominated questionnaires and were therefore biased against male-dominated questionnaires. Sunter's statistical analyses do not identify a preference for female-dominated questionnaires by the multiple evaluation committees. The IRR test results illustrate the majority of the outliers demonstrated a male preference yet, when the final committee evaluations are compared to the consultants' evaluations, the disparities are indicative of a bias against male-dominated jobs.

795. In determining why the differences occur, the Tribunal is entitled to look at some compelling facts. Most importantly, the MEC was conservative in its discipline relative to the consultants. Firstly, according to Willis, the MEC discipline was more accurate than the consultants as reflected in his report to the JUMI Committee (Exhibit R-22) on the re-evaluations of the MEC evaluations which arose out of the IRR 103 challenges and the Treasury Board challenges. That report states in part:

> We have no significant concerns regarding the MEC's understanding and application of the evaluation plan. The MEC's pattern of application of the evaluation plan to positions (their "discipline") differs in some respects from the pattern which the consultants would use. However, given the manner in which the MEC membership was determined, their discipline constitutes a more accurate reflection of the values of positions as commonly understood within the Government of Canada than the consultant could determine from an outside point of view.
>
> (Exhibit R-22, p. 8)

796. Secondly, the Willis consultants had an established discipline prior to the JUMI Study based on their experience in other studies. There is ample evidence from which to conclude the Willis discipline was more liberal than the MEC discipline. According to Owen, another Willis consultant, the Willis discipline influenced the consultants in their

evaluations performed during the JUMI Study. The consultants were experienced and professional evaluators. They were more familiar with higher level jobs, both managerial and technical, which they gained through previous pay equity exercises. The JUMI Study was the first time the consultants had done any evaluations in the Federal Public Service.

797. Thirdly, overall the evaluation committees followed the MEC's discipline. There were three or four occasions where the evaluation committees actually evaluated above the F level to the low G level. According to Willis, this by no means altered the MEC discipline.

798. Fourthly, outliers did not exert an observable influence on the committee evaluations, either in the MEC or in the multiple evaluation committees. The statistical evidence corroborates Willis' own conclusions that the outliers had no discernible effect on the evaluations of the other committee members.

799. Finally, both Willis and the evaluators testified the high level positions were difficult to evaluate. The distribution of questionnaires between male- and female-dominated occupational groups were not the same in terms of value. The more difficult questionnaires were in the high level male-dominated jobs where the greatest difference between the evaluation committees and the consultants occurred.

VIII.CONCLUSION

800. In light of these facts, as well as other matters previously referred to by the Tribunal, it is reasonable to conclude from the conservative discipline established by the MEC, the evaluators' inexperience and difficulty with evaluating high level jobs, together with the very subjective nature of the exercise that the disparity between the consultant and the committee was the result of, and is explainable by those factors we have mentioned. We conclude this resulted in a phenomenon which manifested itself in a reluctance on the part of MEC to attribute high scores to higher level questionnaires. Factors such as weak job information and difficulty in comprehending job information also contributed to this phenomenon. In applying the reasonable standard of proof as required under s. 11 of the Act, it is reasonable to conclude the difference between the committees and the consultants was not a gender difference. We find as a matter of fact the disparities resulted from an inability and/or reluctance on the part of the evaluators to evaluate high level male-dominated jobs according to the discipline of the consultants.

801. The conservative mind set of the MEC evaluators was the origin of this phenomenon which spread and continued throughout the work of the multiple committees. This conservativeness has its most telling effect on the male-dominated jobs at the higher end of the scale.

802. During his testimony Willis was unable to give his unqualified support to the JUMI Study results. He was, however, of the opinion the results should not be "trashed". He was of the opinion they could be accepted at face value or with some adjustments by the Tribunal. There

remained, however, lingering questions, in view of Willis' discomfort, about how well the process worked.

803. This hearing has spanned 232 days to date. The Tribunal was afforded a wide range of both expert opinion evidence and non-expert evidence, including anecdotal evidence. In addressing the issue of reliability, we are mindful of the large number of agreements between the consultants and the evaluation committees on the re-evaluations. The standard of proof in this case is one of reasonableness. We find, for the most part, the committees and the consultants were able to agree on the evaluation scores, except with the more complex, professional and technical jobs distributed at the high range of the male-dominated jobs. The phenomenon beginning with the MEC, carried over into the multiple committee evaluations and was nourished by other factors, which contributed to the disparity between the consultants and the committees.

804. We find as a fact that the evidence establishes the evaluation results are sufficiently reliable, by any reasonable standard, as a basis on which to calculate the existence or otherwise of a wage gap between male and female employees employed in the same establishment who are performing work of equal value within the meaning of s. 11 of the Act and the Guidelines. The Employer has failed to provide any evidence which would cause the Tribunal to find otherwise or to change its decision.

Dated at Vancouver, British Columbia, this 19th day of January, 1996.

Donna Gillis, Chairperson

Norman Fetterly, Member

Joanne Cowan-McGuigan, Member

APPENDIX A
COMMITTEE MANDATES

1. Sub-Committee on a Common Evaluation Plan

(a) Committee Mandate

The official mandate of this sub-committee was to determine what evaluation plans to examine and make recommendations to the JUMI Committee at large.

2. JUMI Committee

(a) Committee Mandate

The task of the JUMI Committee was to develop agreed parameters under which equal pay for work of equal value, as incorporated in the provisions of s.11 of the Canadian Human Rights Act could be implemented and to prepare a detailed plan for its implementation covering that portion of the Public Service for which the Treasury Board represents the employer.

3. Sub-Committee on Communications Strategy

(a) Committee Mandate

The mandate of this sub-committee was to analyze communication alternatives and recommend the most effective ones for implementation.

4. Sub-Committee for Training

(a) Committee Mandate

The mandate for this sub-committee was to draft and recommend a training package for coordinators. This sub-committee later transmuted into the Administrative Sub-Committee.

5. Testing Sub-Committee on the Willis Evaluation Plan

(a) Committee Mandate

The main objective of this sub-committee was to present to the JUMI Committee recommendations related to:

(i) the modification or clarification of the definitions and the factors pertaining to the four evaluation charts of the Evaluation Plan.

(ii) the choice between the Working Conditions Evaluation Chart No. 1 or 2.

6. Sub-Committee on the Willis Questionnaire

A-1

(a) Committee Mandate

The mandate of this sub-committee was to finalize the format and contents of the Willis questionnaire (including developing examples). The sub-committee was asked to review the questionnaire and ensure that the questionnaires were sufficient to gather the necessary data for the questionnaire.

7. Administration Sub-Committee

(a) Committee Mandate

The mandate of this sub-committee was to conduct examination and discussion of, and present recommendations and/or make decisions on, all matters related to the administration of the Equal Pay for Work of Equal Value Study, with the exception of those responsibilities assigned to the Equal Pay Study Secretariat. Specifically, this sub-committee:

(i) devised, implemented and monitored any administrative action required by JUMI;

(ii) provided the EPSS with guidance regarding administrative issues;

(iii)recommended to JUMI actions (to be) taken;

(iv) ensured the smooth administrative operation of the Study, within the framework established by JUMI, through setting priorities, delegating work, resolving issues and assessing the progress of the Study; and

(v) co-ordinated required training to coordinators, evaluators, reviewers and secretaries.

8. Master Evaluation Committee

(a) Committee Mandate

The primary purpose of the Master Evaluation Committee (MEC) was to evaluate a representative sampling of positions and, in so doing, provide the frame of reference for the five evaluation committees (later expanded to nine) to rely on, so that at the conclusion of the position evaluation stage of the study all 4,400 position evaluations would relate to one another fairly and equitably. The mandate of the Master Evaluation Committee was to:

(i) establish benchmark position ratings for approximately 600 positions through initial evaluation of a representative number of positions sampled, and a frame of reference to guide subordinate evaluation committees in the evaluation process;

(ii) provide advise and assistance to subordinate evaluation committees in particularly difficult evaluation cases;

(iii)implement a monitoring system to ensure consistent and bias-free rating by subordinate evaluation committees; and

(iv) as final authority, resolve controversial cases where an evaluation committee has made every effort to arrive at an agreed to rating but has been unsuccessful in doing so.

9. Mini-JUMI Committee

(a) Committee Mandate

The mandate of the Mini-JUMI Committee was to deal with procedural problems arising from the study. Initially, the JUMI Committee dedicated a large amount of time discussing procedural problems but eventually decided to create the Mini-JUMI Committee to deal with them.

10. Equal Pay Study Secretariat

(a) Committee Mandate

The Equal Pay Study Secretariat was a Joint Union/Management Secretariat. It was located in the Jackson Building and provided all administrative support to the evaluation process in the Study. The Chief was responsible for the co-ordination of all support activities and the effective communication of JUMI and Administrative Sub-Committee instructions.

11. Inter-Rater Reliability and Methodology Sub-Committee

(a) Committee Mandate

The mandate of this sub-committee was:

(i) to determine and make recommendations about the methodology and research necessary to test evaluation committee rater reliability; and

(ii) to assess and make recommendations about research methodology as it applies to the JUMI Study as a whole.

12. Five Multiple Evaluation Committees

(a) Committee Mandate

The mandate of the five evaluation committees was to:

(i) evaluate approximately 750 positions each; and

(ii) keep the Master Evaluation Committee abreast of their evaluation proceedings, results and issues, through chairpersons.

The five evaluation committees were reorganized into nine evaluation committees on April 14, 1989.

13. Inter-Committee Reliability Sub-Committee

(a) Committee Mandate

The mandate of this sub-committee was to:

(i) examine the results of the tests administered to the evaluation committees in relation to the baseline provided by the consultants;

(ii) examine the baseline score provided by the consultants;

(iii)determine the significant differences in the consensus ratings of the committees in relation to the benchmarks and the baseline;

(iv) formulate if needed, recommendations for training, re-training by the consultant and/or other courses of action for JUMI considerations; and

(v) identify procedural/process problems and potential for improvement including the revisions to the formulation of rationales.

14. Mini-MEC

(a) Committee Mandate

The Mini-MEC was charged with the task of reviewing the committee challenges to the MEC's evaluations. The JUMI Committee, directed Johanne Labine of PSAC and Michel Cloutier of the Treasury Board, both of whom sat on the MEC, to review the working conditions of all 100 benchmarks for shift work, overtime and living conditions, assess the amount of points to be changed, if any, and correct rationales.

It was ultimately decided that the MEC would not be reconvened, and that a Mini-MEC, a nucleus, or a small number of evaluators from the MEC would undertake this exercise. There were two members from the MEC who were selected to represent this Mini-MEC, Michel Cloutier and Johanne Labine. The idea was that Mr. Willis would meet with the two of them and resolve any differences.

15. Sub-Committee on Total Compensation

(a) Committee Mandate

The draft terms of reference for this sub-committee as of September 21, 1989 were:

(i) To identify the elements of compensation in the Federal Government that comprise wages as defined in Section 11(6) of the Canadian Human Rights Act;

(ii) To compile the data required to establish wages for the positions evaluated;

(iii)To devise a method to cost total compensation for purposes of correcting any identified wage disparities.

16. Quality Analysis Committee

(a) Committee Mandate

Paul Durber of the Commission created the Quality Analysis Committee to examine the 25 male-dominated jobs noted as possibly undervalued in the Wisner report in May, 1990. The purpose of the committee was to shed light on whether the maleness of the jobs, might help to account for their rating and whether, conversely, the differences between Mr. Wisner and the committees were due to simple perceptions of the work.