Rosenberg, Carl Robert

AN ANALYSIS OF THE VALIDITY OF DETAILED OCCUPATIONAL EXPOSURE HISTORIES USING A COHORT OF PCB-EXPOSED WORKERS

City University of New York PH.D. 1983

PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark _✓_.

1. Glossy photographs or pages _____
2. Colored illustrations, paper or print _____
3. Photographs with dark background _____
4. Illustrations are poor copy _____
5. Pages with black marks, not original copy _____
6. Print shows through as there is text on both sides of page _____
7. Indistinct, broken or small print on several pages ✓
8. Print exceeds margin requirements _____
9. Tightly bound copy with print lost in spine _____
10. Computer printout pages with indistinct print _____
11. Page(s) _____ lacking when material received, and not available from school or author.
12. Page(s) _____ seem to be missing in numbering only as text follows.
13. Two pages numbered _____. Text follows.
14. Curling and wrinkled pages _____
15. Other_____

AN ANALYSIS OF THE VALIDITY OF DETAILED OCCUPATIONAL EXPOSURE HISTORIES USING A COHORT OF PCB-EXPOSED WORKERS

by

Carl R. Rosenberg

A dissertation submitted to the Graduate Faculty in Biomedical Sciences in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

1983

This manuscript has been read and accepted for the Graduate Faculty in Biomedical Sciences in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Aug. 25, 1983
date

Chairman of Examining Committee

Aug 25, 1983
date

Executive Officer

Steven B. Blum

John Thornton

Arthur M. Langer
Supervisory Committee

The City University of New York

Abstract

AN ANALYSIS OF THE VALIDITY OF DETAILED OCCUPATIONAL EXPOSURE HISTORIES USING A COHORT OF PCB-EXPOSED WORKERS

by

Carl R. Rosenberg

Co-Advisers: Michael N. Mulvihill, Dr.P.H. and Alf Fischbein, M.D.

A study was undertaken to describe and analyze the validity of self-reported occupational histories obtained from a group of 288 blue-collar workers who had participated in a repeated survey concerning PCB exposure and the prevalence of related health effects at two capacitor-manufacturing plants.

Crude validity was operationally defined as the percent agreement between a worker's self-reported occupational history and an unbiased company record of employment. This score was used in a descriptive phase of the study which was concerned with how inaccuracies in self-reporting were reflective of misclassifications in terms of exposure to PCBs.

In the analytical phase of the study, various independent factors which might influence validity were investigated. Multivariate analysis was employed to examine the effects and interactions of such factors - looking at each factor while adjusting for the effects of all the others.

The results indicated considerable variability in crude validity scores. Misclassification of exposure, produced by the above, could lead to either non-detection of a risk factor or spurious associations in case-control and survey type epidemiologic studies. Validity was significantly influenced by diversity of the job categorical pattern, recall ability, sex, duration of employment, time elapsed between work history events and their subsequent recording, and interviewer objectivity. Many of these factor effects were of an interactive nature. Knowledge of these factor effects would be critical either in planning epidemiologic studies safeguarded against inaccuracies or in determining the efficacy of performing such studies in the first place.

ACKNOWLEDGEMENTS

I wish to express my gratitude to Drs. Michael Mulvihill and Alf Fischbein, my dissertation advisers, and to Drs. Steven Blum, John Thornton, and Arthur Langer, for serving on my committee and assisting me in producing the final draft. I would also like to thank Drs. Irving Selikoff, Director of the Environmental Sciences Laboratory, and Terry Ann Krulwich, Dean of the Graduate School of Biological Sciences, for guidance in orienting my research interests. Additional thanks go to Sidney Sibel and Valerie Josephson for assistance in word processing and graphics. I reserve a special thanks for Dr. and Mrs. Arthur Aufses, Jr. for opening the doors to my research career. Finally, I dedicate this dissertation to my dear parents, grandparents, family, and friends, for without their support, this achievement would never have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION AND LITERATURE REVIEW

INTRODUCTION

There is often a major problem aside from controlling for confounding variables in conducting retrospective epidemiological studies (case-control, cross sectional survey). In such studies the accuracy of the obtained exposure information is questionable, especially when the individuals examined are the sole source of such information. Self-reported information is subject to various uncertainties which can result in misclassification of exposure categories and subsequent erroneous conclusions. The extent to which accuracy would be a concern in any self-reporting study would certainly depend upon the intrinsic reporting ability of the study subjects. However, equally or more important might be extrinsic factors influencing accurate reporting and the type or complexity of the exposure information required by the study design.

Testing self-reported information for validity and possibly reliability represents a way of probing for possible inaccuracies. Validity here is defined as the measure of how closely self-reported information provides an indication of the actual situation. This definition implies that an objective source of information representing the actual situation exists. The fact that the validity procedure requires such an objective source would make it impractical in the study-specific testing of self-reporting accuracy. Logically, if objective sources were available, they would be used in preference to varyingly subjective, self-reported information in determining exposure in epidemiologic studies. Validity may in limited circumstances be used to test study-specific accuracy, e.g. in cases where objective sources become available subse-

quent to the conductance of a self-reporting study or when objective sources are incomplete (i.e sufficient to test validity but not complete enough to determine all aspects of exposure or exposure for all individuals), but its real value would lie in providing information as to why inaccuracies occur and how they could be prevented or minimized in future studies (the main objective of the present study). Reliability is defined here as the measure of consistency of self-reporting on two or more occasions. According to this definition, no objective source is required. Thus, reliability may be of use in estimating accuracy in specific self-reporting studies. The rationale is that the consistency process between repeat and original self-reporting sessions would be somewhat analagous to original report consistency with the actual situation, more so, the greater the interim period between evaluations (indirect accuracy). The degree of this analogy would be the critical factor. The issue of reliability in the present study, for reasons which will become clear later, will be handled by discussion only.

The way in which accuracy of self-reported information is assessed by epidemiologic researchers is directly related to the requirements of the study design. That is to say, the complexity of the data to be tested and the type of problem addressed by the study determine the type and complexity of the particular accuracy-evaluating methodology to be applied.

The data sets analyzed for accuracy in the following literature review are quite varied. Some were collected solely for the purpose of demonstrating accuracy-determination techniques, while others were parts of actual epidemiologic studies where accuracy was a major concern of the investigators. The types of data analyzed for accuracy involve

exposure reports from survey and case-control studies, disease frequency reports used in the formulation of vital statistics and health care planning, risk assessment to subjects and their contacts, health status determination, and others. While the present study is specifically concerned with exposure reporting accuracy in case-control and survey studies, accuracy analysis on any data type serves to illustrate the necessary principles.

SINGLE ITEM VALIDITY

The most basic situation exists in studies where the validity of a single item is in question. Statements, typically yes-no, by patients or respondents about an exposure, attribute, or condition are compared with objective sources of the same information from an authoritative record, diagnostic test, or physical examination. This type of validity consists of two components, namely sensitivity and specificity. Sensitivity is a measure of the percentage of those who truly possess a characteristic and are so classified according to an interview or self-administered questionnaire. Specificity is the percentage of those who do not possess a particular characteristic and are classified as such.

The use of this particular method of determining validity is quite common and standardized in the epidemiologic literature. Therefore, this section of the literature review is restricted to a discussion of some pertinent investigations which relate to the methodology developed in this thesis. Investigations by Dunn and Buell (1), Lerman et al. (2), and Goebel (3) serve to illustrate the use of this method quite satisfactorally.

In the first (1), which is a case-control study, the objective

was to examine the nature of the relationship between cervical cancer in women and the absence of full circumcision of their male sex partners. In doing so, the investigators deemed it necessary to determine the degree of error, if any, in assigning risk or exposure categories and thus ascertain whether or not their computed relative risk figure was subject to bias. To accomplish this, an analysis was performed testing the validity of the circumcision histories given by 166 male subjects. The objective information base was the finding by physicians. A sensitivity of 72.4% and a specificity of 83.2% was found. These figures were based upon the inclusion of both "partial" and "no" circumcision in the same classification (under the hypothesis that only "full" circumcision was protective against cervical cancer). When "partial" and "full" circumcision were combined, sensitivity decreased to only 43.2% while specificity rose moderately to 92.3%.

The investigators' conclusion was that error in classification, i.e. lack of validity, was both evident and serious enough to obscure any real existing difference between cases and controls. They attributed this lack of validity to interviewer bias (Interviewers were aware of the identities of cases and controls.), lack of knowledge about circumcision on the part of non-Jewish men (exacerbated by impaired recall in older non-Jewish men), and the fact that the circumcision operation did not produce uniform results (full vs. partial). What was particularly striking in reviewing this study was the fact that what appeared to be a simple, non-recall influenced, task on the part of the respondents - the identification of a singular, seemingly obvious anatomical entity - lead to such a high misclassification rate.

The second study (2) aimed at testing the validity of a past me-

dical history of rubella. It was noted that 20% of women of childbearing age were at risk for rubella during pregnancy at that time. Taking a history for previous rubella had been suggested as an efficient and relatively inexpensive method of identifying women at risk. The validity base was a test for rubella hemagglutination (immunity) in each subject. Of 275 women, 159 gave a positive history of rubella - It was this subgroup that was evaluated for validity. The sensitivity of the rubella history was relatively low, 68.6%. Of all the 121 subjects who had a positive immunological test for rubella, 38 denied past rubella. Specificity also was low, only 82.5%. Of 38 women with no detectable immunity, 7 claimed past rubella.

The conclusion by the author was that the self-reported rubella history was invalid. He attributed non-validity to the fact that the clinical diagnosis of rubella was difficult, easy to confuse with other skin rashes, and sometimes occurred without a rash or as a subclinical disease. Surprisingly, there was no correlation between number of years elapsed since infection and accuracy of self-reporting. Unlike in the previous study (1), non-validity in this case, although also only involving a singular item, was more understandable. The item in question here was neither as evident nor was it in all cases recognizable. Influencing validity, in effect, were both reporting ability of the respondent and the extrinsic factor of subclinical infection. It should be understood, though, that the source of the non-validity, intrinsic or extrinsic, may be irrelavent in terms of practical effectiveness of the rubella history.

Goebel (3) studied the problem of patients transmitting hepatitis B to their dentists. In an attempt to identify such individuals, a

validity study, misdefined by the author as a reliability study, on self-reported medical histories was conducted. A total of 272 individuals was screened. Their self-reported histories, concerning previous exposure to hepatitis virus, were compared to blood tests probing for specific antibodies, the validity base being serologic evidence of hepatitis B. Of the 4 individuals who had serologic evidence of hepatitis B only 2 (50%) reported disease history, jaundice, or illicit drug use. This suggested that hepatitis B carriers may not have been adequately identified by a medical history. Although this 50% sensitivity figure was based on a small sample, it did illustrate the method of single item validity and could possibly have indicated an actual real difficulty in identifying hepatitis B carriers. This was important in that hepatitis B is a serious disease and thus could have serious implications in terms of its effect on dental care delivery.

As with the rubella history study (2), non-validity in this study also had a basis in the extrinsic factor of non-recognizable subclinical infection. Once again, however, the reason for non-validity would make no difference to dental practitioners who for whatever reason would still be at risk for hepatitis B infection.

MULTIFACTOR VALIDITY - BASIC APPLICATIONS

The studies just discussed (1,2,3) addressed the issue of single item validity, each employing the same simple standardized technique. When more than one factor is involved, i.e. items with dimension(s), strings of events, or histories, however, validity determination becomes more difficult. There is no standardized procedure in such cases because the data sets can exist in a variety of formats and with different

levels of complexity. Researchers have therefore formulated specific methodologies for use in such analyses with varying degrees of complexity.

A very basic way of looking at multiple item validity is illustrated in studies by Madow (4); the Commission on Chronic Illness, CCI (5); and the U.S. National Health Survey, USNHS (6); which were concerned with the accuracy of chronic disease prevalence data generated from interview-solicited self-reported information. In each study, at the end of a 12 month period, patients' statements concerning chronic conditions were compared with information compiled in medical records. Specifically, the percentage of matching between the two sources was the validity-determination technique. The study populations of Madow (4) and the USNHS (6) were drawn from health insurance plan groups and consisted of 6,000 and 3,937 subjects, respectively. The CCI (5) study group was composed of 809 inner city residents.

In Madow's study (4), only 54.7% of chronic conditions listed in the medical charts were recorded during the subsequent personal interviews. The figure for the CCI study (5) was even lower, 22.2% (29.6% excluding conditions about which the subject could not have known). In the USNHS (6) study, chronic conditions were divided into three classes: 1) Class I - those appearing on the standard NHS checklist, 2) Class II - those close to or related to those on the checklist, and 3) Class III - non-checklist conditions. The breakdown of conditions by class was 40%, 25% and 34%, respectively. The respective interview-reported percentages were 44.1%, 27.6%, and 20.4%.

It would have been unfair to compare the percent reporting figures from each study in terms of which represented higher or lower

validity in that the population bases were different, these figures were subject to the effects of various confounders, and the study designs differed in some respects (including quality). Madow's study (4) was the least sophisticated and among other things suffered from the problem of uncertainty as to the definition and categorizing of chronic conditions. This may have led to the overlooking of potential matches in cases where the subject described a particular condition in terms different from that of the physician. The CCI (5) and USNHS (6) avoided this pitfall by setting up unambiguous matching criteria based upon the international ISC-PHS code. Both Madow (4) and the CCI (5) used standard medical charts as the base of validity. On the other hand, the USNHS (6) opted for the Med-10 form. The latter insured a more objective validity base, in that in contrast to the standard chart, it provided no detailed medical history, evaluation of symptoms, or differential diagnosis, but merely listed a terse diagnosis for each clinic visit. At any rate, though, none of the investigators considered their percent reporting figures to represent adequate validity.

The low percent reporting figures were attributed mainly to short-term memory effects, lack of knowledge about various conditions, and non-cooperativity. Considering short-term memory effects, Madow (4) found that those subjects whose last clinic visit was closer in time to the interview reported a higher percentage of conditions. Similarly, the USNHS (6) discovered that the frequency of clinic visits was correlated with higher percentage reporting. The implication is that clinic visits may have increased or reinforced awareness of chronic conditions which counteracted an existing short-term memory effect. Lack of knowledge by the subjects about certain conditions, the CCI (5) found, re-

duced percent reporting by 7.4. Therefore, they claimed that the "corrected" figure was really 29.6% (22.2% plus 7.4%). However, the actual figure of 22.2% was the more realistic of the two in that regardless of what caused underreporting, any underestimate of prevalence rates arising from it would have been determined by the actual figure (analagous to the effect of subclinical infections on single item validity). Cooperativity, or attitude of the respondent, defined by Madow (4) as the willingness to allow Census Bureau access to medical records, affected the percentage of reported conditions. Those who permitted access reported 56% of conditions while those who did not only reported 38%. In the USNHS study (6), a similar tendency was found.

Demographic factors, as a rule, appeared to have little impact on percent reporting. One exception was age - Madow (4) found that those who were over 65 reported a greater percentage of conditions (63.5%), while the USNHS (6) found a positive age-percent reporting relationship. This age effect was attributed to increased prevalence of chronic conditions in higher age groups which resulted in a greater awareness of them and increased clinic visits. The CCI (5) in another exception found that women reported slightly better than men (32% vs. 26%) and whites outreported blacks by 33% to 25%. No significant effects of educational level or family income were found in any of the studies.

Additional factors affecting percent reporting were the nature and seriousness of the condition(s), interviewer bias, and interview by proxy. In the USNHS study (6), the first two of these three factors were deemed responsible for the greater rate of reporting of Class I conditions. Interviewers were found to have inquired more intensely about conditions on the checklist and checklist conditions tended to be of a

more serious nature (also leading to more clinic visits). Concerning interview by proxy, self-respondents reported Class I conditions at a rate 6.7 percentage points higher than proxies.

Overreporting error, i.e. conditions reported in the interview but not clinically evaluated, was not considered by Madow (4) but was by the CCI (5) and USNHS (6). It was found by the CCI (5) that 29.7% of conditions listed in the interview were not confirmed by diagnostic evaluation. The figure was 39.6% (unmatched to the Med-10s) in the USNHS study (6). Overreporting was largely accounted for by familiar or self-diagnosed conditions which did not demand medical attention. As a final exercise by the CCI (5), prevalence rates for various diseases were calculated using both interview and diagnosis data. As expected, there were considerable differences, predominantly underestimates, due to self-reporting errors.

The methodology demonstrated in the last three studies represented a first step in bridging the gap between single item and multifactor validity. Instead of investigating one particular characteristic, attribute, or condition, a collection of such was observed and analyzed. This change from single to multifactor validity created two major difficulties. First, the standard sensitivity-specificity measure was no longer applicable. A new technique therefore had to be developed to evaluate validity. Replacing sensitivity and specificity was the frequency of matching (of chronic conditions) between medical records and personal interviews. Using this technique, lack of validity could have been defined in terms of underreporting and overreporting. However, in all three studies only underreporting error was stressed. This may have been adequate for the purposes of evaluating chronic condition reporting

(Most important chronic diseases do seem to be underestimated in terms of prevalence) but perhaps not in other situations, such as in cases involving cumulative workplace exposure. As will be shown later, more refined techniques, combining the two types of error into one measure, have been developed. The second difficulty was that the validity base could no longer be composed of a singular examination or diagnostic test. Instead, medical records, sometimes in combination with the results of comprehensive physical examinations, were used. In the single item case, lack of objectivity and uniformity of the validity base is not a major potential problem. With records, charts, and examinations, the situation is entirely different. Multiple entries on records and charts recorded at varying times and to varying degrees would tend to result in less uniformity and objectivity. Maintaining records and conducting comprehensive physical examinations also requires the participation of more individuals, introducing the effects of interobserver differences and varying interpretations. These problems were quite evident in the studies by Madow (4) and the CCI (5). In the USNHS study (6), they were recognized and attenuated somewhat (Med-10 forms). Multifactor validity determination also presented similar type problems for the interviewing process. Once again, the USNHS study (6) attempted to take them into account (3 classes of chronic conditions).

In terms of comparison with other self-reporting studies and methodologies, validity as determined in the studies just discussed (4,5,6) did not involve any long-term memory effects. The subjects were only required to list existing or recent chronic conditions and did not have to recall relatively distant past events, place them in time, report dimensions of the events, nor piece them together in the order in

which they occurred. In other situations, such as those inolving exposure to harmful substances in the workplace, the methodology of validity determination would have to take memory factors into account. In addition, validity, as determined here and in virtually all studies to date, was only representative of a group, not of any individual.

MULTIFACTOR VALIDITY (OR RELIABILITY) - ALTERNATIVE BASIC APPLICATIONS

In the following set of papers, the methodology of validity determination is basically similar to that discussed previously. However, here it is applied to varying situations, with various refinements, and with adaptations to these new situations. Additionally, the parameter of reliability is introduced.

As an introduction to this section, a discussion of an unsuccessful attempt at a validity analysis of the type at issue here is in order. Petiti et al. (7) designed a study whose objective was to determine the validity of a self-administered questionnaire. The study sample consisted of 267 volunteer subjects (aged 18 to 72; 176 females, 91 males) who were members of a health insurance plan undergoing regular checkups. Each individual completed two questionnaires - one on smoking habits and the other on health status (which included questions on smoking). To serve as a validity base, blood samples were taken, but the subjects were not informed as to their purpose. These blood samples were analyzed for levels of SCN and CO, specific physiological indicators of smoking. Unfortunately, the blood levels of these substances were found not to be sensitive enough indicators of smoking to serve as an adequate validity base. Thus, the only determination the authors were able to make was that there was 95.4% consistency in responses to

questions concerning smoking for the two different questionnaires. This result introduced the issue of reliability which will be dealt with later in this section. In conclusion, this study emphasized the importance of a good validity base. The fact that the authors recognized the deficiency in their method, before erroneous conclusions could have been formulated, effectively brought this point across.

In the following paper, by Brady and Martinoff (8), validity was defined as the percentage of a group of individuals that was consistent/inconsistent in an interview or questionnaire vs. record comparison. This methodology of validity determination was not much different from that used in the most recently discussed group of papers (4,5,6), with the exception that validity was defined more directly in terms of people rather than one of their attributes or characteristics. As in the study by Goebel (3), the object of the investigation was the validity of medical history data obtained from dental patients. In this case, however, the purpose was not only to use validity as a means to an end (i.e. determine if self-reported data was adequate for disease detection) but also to look specifically at validity itself and factors which could have influenced it. The health history was obtained from 2,107 patients, all of whom completed the forms either at home or in the dental office, without the assistance of any interviewers. These patients were neither screened nor selected but were admitted to the study as they applied for treatment at a university dental center. The validity base consisted of a physical examination augmented by a history taken by student dentist interviewers. This history included discussion focusing upon unanswered questions and inconsistent answers. Following the physical examination, each subject was placed in one of four health catego-

ries. The idea was to observe how or if health status influenced validity.

According to the described procedures, only 68% of the patients gave questionnaire data that were completely valid. Inaccurate or incomplete information, as compared to interview followup and physical examination, was provided by the other 32%. No relationship between validity and health status was found. The investigators concluded that the questionnaire information could not be assumed to be valid and that dentists therefore could not rely upon patients' answers to health questions before initiating treatment. Possible reasons for non-validity were: 1) a "no" answer may have either meant that the patient did not understand the question or that the symptoms or disease never existed; 2) a question may not have been taken seriously since its relationship to dental complaints may not have been perceived, 3) stress, 4) and the wording of questions. Shortcomings of the validity base may also have affected the results. The problems with uniformity and objectivity in conducting physical examinations were discussed in the previous section. Interviewer error, absent from the questionnaire phase, may have been introduced in the validity-determination stage. The authors noted that although each student interviewer was given standardized instructions and evaluated by faculty members, they were still inexperienced. At any rate, some of the information they gathered was influenced by the questionnaire answers.

Using the above methodology, validity was once again a function of a group. Since validity was defined as the percentage of a group that gave a totally correct history, individual degrees of validity were not considered. One or any number of inconsistencies between the question-

naire and the validity base would have placed the subject in the invalid group (32%). If all 2,107 subjects had given only one incorrect answer, the validity would have been 0%. Perhaps in this case, even one error may have represented a serious enough problem to define it as non-validity. If not, a weighing system may have been desirable.

The methodology employed by Sacks et al. (9), in the next paper, was quite similar to that just discussed, the only difference being that reliability rather than validity was the subject of investigation. A general definition of reliability was presented at the outset of this study, but specifically, in this paper, it was defined as the percentage of a group of individuals consistent/inconsistent in a test/retest situation. Investigated here was the reliability of the health hazard appraisal (HHA) questionnaire which is used by physicians to help achieve risk reduction in individual patients. A sample of 207 subjects was drawn for this reliability analysis. Informed consent was obtained. The study population consisted of four respondent types: physicians, nurses, federal employees, and patients. Each subject first filled out a baseline HHA questionnaire. An average of 85 days later, 203 out of the original 207 then filled out a followup. The 85 day gap was instituted to prevent answers on the first questionnaire from influencing those on the second. However, 85 days represented a sufficiently long period for certain self-reported behaviors, such as smoking, to change. If such changes did occur, one would not have obtained a true measurement of reliability. To prevent such confounding effects, the HHA questionnaire only addressed items which should have been constant. This issue of selection of gap length in reliability studies is quite crucial as will be seen in several subsequent studies.

As an overall observation, only 15% of the subjects had no inconsistencies when comparing original and followup questionnaires. This figure did not vary significantly among the four respondent gropus. In terms of specific questions, the greatest number of inconsistencies occurred for the question concerning "number of miles driven in the past year." Six out of ten respondents (66% of males, 51% of females) changed their original answers by an average absolute value of 4,700 miles/year. The overall tendency was to report an increase on followup. An absolute height difference of 1 inch or greater was reported on followup by 9% of the respondents (11% of males, 7% of females). Parental age was changed by 33% of the respondents (36% of males, 29% of females). One in five ex-smokers had difficulty in consistently recalling average consumption (27% of males, 9% of females). The same was true of "number of years since quit smoking" (18% of males, 22% of females).

These results, for the most part, seemed to indicate a greater consistency in female reporting. One exception was alcohol consumption by ex-drinkers. Contradictory responses were given by 33% of the respondents (30% of males, 40% of females). This may have been explained by an increased social stigma attached to women's drinking problems. For females only, 17% were inconsistent about Pap-smear history and 13% about age at first intercourse.

The question of group percentage versus individual degree (of reliability in this case) arose here as it did in the previous validity study (8) using the same methodology. However, for the HHA questionnaire, it appeared that an incorrect answer to just one question could have been crucial. Data from such a questionnaire were used to measure overall risk. Changes in height, weight, and parental age could have

affected risk multipliers for obesity and heart disease by 0.1 and 0.2, respectively. Since heart disease was a major and frequent cause of death in the older age groups, a small change in risk multiplier could have caused a great change overall risk estimate. The conclusion was that the reliability of the HHA questionnaire was low.

MULTIFACTOR VALIDITY (OR RELIABILITY) - INCREASED DATA SET COMPLEXITY I

In the following two papers, more complex issues, and hence methodologies, were applied to validity determination. In the first paper, by Corwin et al. (10), the effects of time lapse and recall ability on validity were investigated. Their study involved 90 patients presently or previously hospitalized for peptic ulcers. The demographic breakdown was: 50 white-40 black and 80 male-10 female. The mean age was 52 years. The validation procedure involved the comparison of interview and hospital record data using Pearson's correlation coefficent (r), supplemented by group consistency and matching percentages. Twenty-five patients were interviewed while still hospitalized, with the remaining 65 being interviewed post-discharge, concerning past hospitalizations for peptic ulcer. The post-discharge group was broken down into four subgroups based upon elapsed time since hospitalization: 1) less than 6 months - 10 patients, 2) 7 to 18 months - 21 patients, 3) 19 to 60 months - 19 patients, and 4) more than 60 months - 15 patients. The interviews were conducted as if a routine medical history so that the patients were unaware a study was being conducted. The pertinent validity base information consisted of date of hospitalization, weight at admission, and ulcer-related symptoms.

The majority of the group of 65 former patients recalled to with-

in 3 months their dates of hospitalization, but 6 (9.1%) made errors greater than ± 1.5 years. The accuracy of admission date correlated positively with the recentness of hospitalization ($r = .71$, $p < .01$). There was considerably greater inaccuracy in reporting weight at admission, again, more so with longer time elapsed since hospitalization. The mean weight error was ± 3 pounds for those presently hospitalized but ± 11 pounds for the "more than 60 months elapsed time" subgroup. From this last observation one could surmise that the mean error figure for this particular subgroup consisted of a ± 8 pound recall component in addition to the ± 3 pound baseline error attributable to imprecision (ignoring the possibly confounding effect of intentional misreporting). Accuracy in weight reporting was also correlated positively with recentness of hospitalization ($r = .89$, $p < .01$). There was generally good agreement between interview and records in terms of potential underreporting of ulcer symptoms. For example, 51 of 57 (89.5%) post-hospitalized patients who had the presence of pain listed on their charts reported this during interview. In comparison, 24 of 24 (100%) of the presently hospitalized group did the same. However, overreporting error was substantial. For example, 11 of 40 (27.5%) post-hospitalized patients reported tarry stools in the interview which was not verified by the records. The overreporting rate for this symptom was 4 of 17 (23.5%) in the presently hospitalized group. Overall, note the recall error effect, i.e. lower accurate reporting for post-hospitalized group. The accuracy of symptom reporting was positively correlated with the recentness of hospitalization ($r = .66$, $p < .01$).

In a second paper, by Chamberlain and Johnstone (11), the validity of pregnancy histories was investigated. The specific objective was

to test womens' recall abilities concerning events in previous pregnancies. The validity base, hospital pregnancy records, was considered to be of high standard. The study group consisted of 174 multiparous patients. They all had had at least one previous delivery at the particular hospital being investigated, with full records available for each respondent. During their current pregnancies they were asked about their obstetric histories with specific reference to: 1) date of birth for each child, 2) lengths of corresponding labor periods, and 3) birthweight for each child.

Dates of birth were accurately reported for virtually the entire sample, with one exception for a child born several minutes before midnight. For birthweights, 73% of the sample were accurate for all children within ± 4 ounces. The figure was 77% (72 of 94) for the penultimate pregnancy, while for all previous pregnancies it dropped to 68% (55 of 80). Overestimates occurred for 11% (19 of 174) of the women with underestimates occurring for 16% (28 of 172). Length of labor, a more difficult variable, was not recalled as accurately. Only 52% (90 of 172) were accurate within ± 4 hours. The accuracy figures for penultimate and all earlier deliveries, respectively, were 73% and 57%. For all errors, the tendency to overestimate was nearly twice as great as that to underestimate. The lower validity figures for earlier pregnancies indicated that memory played a major role in the validity of the pregnancy history. One should have noted the relatively lower accuracy for length of labor which was not merely an event but one with the dimension of duration. No data were presented concerning validity differences between those women with differing numbers of deliveries. Number of deliveries may very well have interacted with the time-lapse effect.

In the overall scheme, the two papers just discussed (10,11) demonstrated the increased level of difficulty that would be faced by investigators who must rely upon data subject to time-lapse, i.e. long-term memory, errors. Up until this point, non-validity (or non-reliability) was mainly a function of short-term or time-independent error such as that attributable to short-term memory lapse, oversight, lack of knowledge, misinterpretation, non-cooperativity, and interviewer bias. In self-reporting, retrospective studies, especially those requiring years of recall on the part of the respondents, non-validity due to memory loss would become an additional and, depending on the length of the recall period, a critical factor. It is also possible that time-lapse may interact with, for example, the complexity of the recall pattern, i.e. the number of items and accompanying dimensions.

MULTIFACTOR VALIDITY (OR RELIABILITY) - INCREASED DATA SET COMPLEXITY II

In the following section, more statistically refined methodologies of validity (or reliability) determination were employed, either to handle more complex data sets, more thoroughly analyze simpler data sets, or both. In one subgroup of three papers, the major technique of accuracy measurement was the correlation coefficient. In the first of these papers, Norell (12) was interested in determining through personal interviews the rate of noncompliance in patients who were instructed to take the drug pilocarpine, at specified times, for the treatment of glaucoma. A sample of 82 outpatients at an eye clinic, all meeting specific diagnostic criteria, was instructed to take 4% pilocarpine eyedrops, three times daily, for a 20-day period. The sex breakdown was: 45 males and 37 females. The median age of the patients was 73

years. The interview results were compared with those from an objective monitoring device which recorded self-medication (validity test). At the end of the 20-day period, 73 patients of the original sample were interviewed concerning non-compliance over the last 7 days.

The data were first analyzed in terms of sensitivity and specificity. The sensitivity was extremely low - Only 7 of 36 patients (19.4%) who missed one or more doses reported this fact when interviewed. Not surprisingly, those who missed no doses reported this fact quite accurately. The specificity was 91.9%. To further analyze the data, the Spearman rank correlation coefficient was calculated. This procedure involved the separate ranking of monitor and interview data, with a correlation coefficient being determined from these ranks. The result was a statistically significant but only small positive correlation between the two sources ($r_s = .38$, $p < .001$). In interpreting this figure, it should be understood that statistical significance only indicated that the correlation coefficient was greater than zero and not that the interview was valid. Indeed, few researchers would consider .38 correlation as indicating practical validity. As such, the r_s value was consistent with the low sensitivity figure of 19.4%.

The author's conclusion was that the interviews were not valid. One possible cause for this low validity was the unwillingness of the patients to admit non-compliance. For example, a particular patient who thought his or her non-compliance was unusually high might have tended to underestimate this fact during the interview. The above may have been partially due to the relationship between interviewers and patients, i.e. higher validity may have been obtained by using interviewers outside the staff, not known personally by the patients. Secondly,

memory loss may have played a role - There may have been difficulties in recalling the number of missed doses over a relatively long time period, especially since what was being asked for was reporting of a negative event, i.e. something that didn't happen.

In a second paper, Andrasik and Holroyd (13) were concerned with the ability of questionnaire data to provide accurate estimates of headache frequency, intensity, duration, and symptomatology. They noted previous headache questionnaire studies which demonstrated high test/retest reliability. However, they were not convinced of the validity of such questionnaires. They were, in essence, inquiring if reliability predicted validity. In their study, questionnaire data were compared with daily chart recordings by the participating subjects (the base of validity) in terms of headache frequency, intensity, duration, and symptomatology. No interviewers were used. The subjects were divided into two groups: 1) 33 out-patients with a mean age of 35.2 years who had been undergoing headache treatment for the past ten years and 2) college students who suffered from tension headaches. The latter group was split into two subgroups, one of which filled out the same questionnaire 2 weeks later (for a reliability evaluation). The daily recording process was conducted during the 2-week interim period defined by the two questionnaire administrations. For both validity and reliability analysis, the methods of comparison were Pearson's correlation coefficient (r) and Student's t-test.

Test-retest reliability was considered high by the authors. Pearson's r values for headache frequency, intensity, and duration were .84, .66, and .77, respectively ($p < .0001$). In terms of practical reliability, though, only the frequency value appeared to be marginally

convincing. Note that the interim period was relatively short. Indeed, the group mean comparisons uncovered significantly higher intensities (Student's t-test, $p < .04$) and lower durations (Student's t-test, $p < .05$) on the repeat questionnaire.

The validity test results, unlike those for reliability, indicated only minimal correspondence between the initial questionnaire and recorded data. Of a total of eight variable comparisons, only two Pearson r values were significantly different from zero. Furthermore, neither of the two statistically significant values (.39 and .71, $p < .05$) were convincing in terms of practical validity. Looking at group means, the tendency was for all subjects to underestimate frequency on the questionnaire, but this difference only reached significance for the retested subgroup of college students (Student's t-test, $p < .0001$). All subjects reported significantly greater pain intensity on the questionnaire ($p < .0001$).

The authors concluded that the questionnaires <u>may</u> not have provided a valid index of headache activity in spite differing reliability indications here and in other studies. It could well have been that the low validity was due to the complex (multifactor, multidimensional) nature of the information required by the study design. In interpreting the results, though, two aspects of the methodology had to be considered. First, while the daily recordings satisfied one major criteria of a validity base (no recall bias effect) they were compiled by the subjects themselves and thus vulnerable to other error sources. Note especially that they were compiled under stressful conditions, i.e. during headaches. Secondly, there is a major pitfall concerning the use of Pearson's r as a test of consistency which will be described after analysis

of this subsection.

In a third paper involving evaluation by correlation coefficient, Streissguth et al. (14) investigated the reliability of self-reported drinking behavior. Specifically, they tested if an individual classified as a "heavy drinker" or a "light drinker" on the basis of a first interview would still be classified as such upon reinterview approximately one week later. The total sample under investigation consisted of 78 pregnant women, all in their fifth month. The first 67 were selected consecutively from three prenatal clinics, but since few reported high alcohol consumption, 11 heavier drinkers (3 or more drinks per day) were added. The women were predominantly white, married, and middle class. Their mean age was 26 years with a range of 15 to 41 years. The rationale for investigating drinking behavior in pregnant women was the damage alcohol could do to unborn children. Derived from each interview were three alcohol consumption scores: 1) AA - absolute alcohol consumption, 2) QFV - quantity frequency variability index of consumption, and 3) VV - volume variability index of consumption. Two interviewers were employed for this study, but each subject got the same interviewer for test and retest. Filler questions were included to help prevent the original answers from influencing those on the retest. Each interview was conducted at home and covered two subperiods: the 5 months since onset of pregnancy and the month preceeding pregnancy. The answers to questions were forced choice.

The AA score reliability was tested using Pearson's r. The test-retest correlation coefficients were .90 and .89 for pregnancy and pre-pregnancy subperiods, respectively. Heavy drinkers, specifically, showed more variability in AA scores. This could have been due to two

factors: a greater number of possible replies to interview questions and the mind altering effects of heavy alcohol consumption. AFV scores were determined for wine, beer, and liquor, with reliability being tested in this case by Kendall's τ rank correlation. The results for wine, beer, and liquor were .90, .90, and .84, respectively (.85, .86, and .85, respectively, for pre-pregnancy) - similar to those for AA scores. VV scores (11 non-ordered categories in massing and spacing of drinking) were tested for reliability by measuring the proportion of consistency, i.e. the number of subjects in the same category upon retest. For pregnancy and pre-pregnancy the proportions were .68 and .63, respectively.

The authors considered test-retest correlation or consistency to be high enough to allow reliance upon the self-reports of drinking behavior in pregnant women. They did admit that some subjects showed considerable variability on retest which was not taken into account when defining reliability or validity by a group figure (weakness of the correlation coefficient and all other methodologies evaluated up to this point). They went on further to state that the inclusion of filler questions in the interview made the subjects feel that their drinking habits were not being singled out for censure, yielding increased cooperativity. As a final note, one might take exception to the VV proportions as indicating high reliability.

Some general comments about the subgroup of three papers just described (12,13,14) are now in order. The use of the correlation coefficient may add statistical refinement to the measure of validity and/or reliability but its utility has the potential of being overestimated, overextended, and perhaps misinterpreted. In the first two papers (12, 13) statistically significant but relatively low correlation coeffi-

cients could have been taken to represent high or acceptable validity or reliability. One study, not described in this section, by Thompson and Collins (15) did just that. For example, they reported a Pearson r value of .48 ($p < .001$) for headache severity to indicate consistent reliable reporting. Objectively speaking, it would not seem that an r value of .48, no matter how statistically significant, could have been interpreted as such. The error in this case was the implication that statistical significance meant practical significance. Given that the same subjects were interviewed twice to generate these correlation data, it was no surprise that the r value had a low probability of having occurred by chance. The lesson to be learned was that practical reliability or validity demands more than just a low probability of chance occurrence. Situations may also arise in which the r value can appear to contradict group mean or percent agreement results. As it turns out, there would not necessarily be a contadiction. High r values could be compatible, for instance, with relatively lower percent agreement in cases where most or all the subjects tend to overestimate or underestimate particular levels of occurrences (i.e. undirectional error). Thus, high r values do not necessarily mean high reliability or validity. They must therefore be interpreted in light of the specific circumstances of the study in which they are employed. A prime example of their overextension will be seen shortly.

A majority of the papers discussed up to this point have made use of the percent agreement technique to measure validity and/or reliability. The analysis of binary response questions, typically yes-no, is one specific case in point. Sanders (16), in a critical review of morbidity surveys, provided an example of yes-no analysis. A sample of

106 families, representing 377 individuals who had been previously interviewed at home, received a physical examination at a local health clinic approximately one week later. The identical yes-no checklist of 32 chronic conditions was used in both the interviews and physical examinations. In only 33% of the response pairs was the reply given to the physician the same as that given to the interviewer. In 57%, the response to the physician was "yes" while the corresponding response to the interviewer was "no." The converse situation occurred for the remaining 10%. One explanation of the 67% non-agreement was interview by proxy (Non-agreement was 74% in cases where mothers responded for children.). However, adjusting for this only reduced the figure to 62%. The relatively small effect of proxy-interview was previously cited in the discussion of the study by the USNHS (6). Any male-female differences were not noted. One postulated reason for the non-agreement was a psychological factor, such as the belief concerning possible benefit or harm that might have come to the patient as the result of a particular reply. For this reason, the respondents may have been more motivated to give a correct reply at the clinical examination than at the interview.

In some of the studies discussed up to this point there probably existed significant confounding (e.g. bias, design errors) which tended to make validity or reliability appear higher or lower than the actual value. In other cases the methodology of evaluation was accurate, but accuracy could have been increased with question refinement, memory stimulation, or interviewer training. However, specifically in yes-no situations, their exists a factor working in one direction, tending to inflate actual validity or reliability figures. This factor is chance agreement, i.e. agreement not attributable to accuracy. It is easy to

see how chance agreement can occur in a binary response system which is the equivalent of guessing on a true-false test. This chance agreement potential would vary according to the relative numbers of "yes" and "no" replies.

Recognizing this, researchers have developed several correction factors in an attempt to arrive at a truer figure. A study by Paganini-Hill and Ross (17) addressed both this issue and one discussed earlier - the limitations of Pearson's r in assessing accuracy. These investigators noted that since much epidemiologic data are obtained through personal interviews, the quality of these data depends upon the ability of study subjects to report information accurately. They emphasized that critics have often questioned the validity of case-control studies on this basis. Consequently, in the context of a case-control study investigating the relationship between menopausal estrogen therapy and breast cancer, they proceeded to compare self-reported exposure information from their subjects with available records containing some health-related information about them.

A total of 393 women, residing in two predominantly white, upper middle-class retirement communities (131 cases, 262 matched controls), comprised a potential validity-determination group. Their mean age was 71 years with a range from 57 to 79 years. All cases and controls were to be assessed for non-validity (possibility of misclassification), but medical records, the main base of validity, were only available for 334 of the subjects. Questionnaires concerning exposure were administered to the 334 women by one nurse-epidemiologist interviewer, eliminating the chance for inter-interviewer bias. Pertinent questions focused on drug usage; menstrual, reproductive, and physical characteristics; and

histories of certain diseases. Medical record abstraction was done by one physician-epidemiologist and included information similar in type to that gathered by personal interview. For the drug usage questions, pharmacy records were also available. One must note again the possible incompleteness of and/or errors in medical records, concerning their use as a validity base. Agreement between the two sources of data was measured by the K-statistic and Pearson's r. The former was used for binary response analysis while the latter was used to analyze continuous data. The K-statistic formula was as follows:

$$K = (A_{obs} - A_{exp})/A_{exp}$$

where A_{obs} = observed (raw) agreement and A_{exp} = the proportion of agreement expected by chance as calculated from 2x2 contigency tables. K-values may range from +1.00 (complete agreement) to -1.00 (complete disagreement) with a zero value corresponding to agreement predicted by chance alone.

Pearson correlation coefficients (interview vs. medical record) for particular variables were .87 (height), .89 (weight), .78 (age at last menstruation), .95 (age at hysterectomy), and .96 (number of children). In terms of raw agreement, 87% of the subjects reported height within 1 inch, with a slight tendency to overestimate during the interview (close agreement with r value). However, only 71% of the weight measurements agreed within 10 pounds. There was a strong tendency to underestimate this variable during the interview. This was a perfect example of the failure of the r value to measure validity in cases of unidirectional over- or underestimation (i.e. The r value of .89 was not

a sensitive indicator of weight agreement). Specifically, many of the women understated their weights to a similar degree (intentionally or subconsciously). The similarity of the discrepancies produced the high r value while the discrepancies themselves accounted for the lower raw agreement figure. It is assumed that the r-value of .89 for weight was highly statisically significant, but it certainly did not indicate practical validity. Raw agreement and correlation was lower for age at last menstruation. Only 68% reported this age within 1 year. Long-term memory lapse was a factor in this case. The correlation coefficient for age at hysterectomy was high but, as with weight, disagreed with the raw agreement percentages - 81% agreed within 3 years but only 68% did so within 1 year. Seven subjects did not agree on the fact of a hysterectomy. Long-term memory lapse was in effect here too. Agreement on number of children matched exactly for 96%, consistent with the r value of .96.

Moving to the binary response results, agreement for history of selected diseases ranged from a low of 87% (κ = .63) for benign breast disease to 96% (κ = .70) for diabetes. Agreement on use of selected drugs ranged from a low 69% (κ = .38) for barbituates to a high of 87% (κ = .60) for antihypertensives. The percent agreement figures for drug usage depended on drug type. Confusion over names and multiple drug uses contributed to decreased validities. Agreement (using medical charts) for oral estrogen replacement (ever/never) was 75% (κ = .51). The Pearson correlation coefficient for "number of months used" was .63. Using pharmacy charts, agreement for oral estrogen replacement dropped to 57% (κ = .21) and the correlation coefficient for "number of months used" to .30. Agreement between medical and pharmacy charts was only

74% (κ = .42) with r = .32 for "number of months used", indicating a possible weakness in the validity base (Medical records were only initiated at the time of entry into the community.).

Overall, the authors considered validity to be satisfactory. This assessment seemed to have stemmed from the fact that cases and controls showed no validity differences. However, the ability to detect any exposure effect might very well have been impaired if one looks at some of the relatively low κ-values for the binary responses (insofar as κ is a satisfactory correction factor). Note that the κ-levels were considerably lower than the corresponding raw percent figures. As a consequence, one should at times be wary of raw percent figures (e.g. in some of the papers analyzed up to this point). The fact that several correlation coefficients tended to misrepresent validity (i.e. differed from raw percentages) on account of unidirectional misestimate was not critical in this study since cases and controls were equally affected. However, in other situations where cases and controls were not similar (e.g. unidirectional misestimate, but to a differing degree or in a differing direction between the two groups), high r values could coincide with the existence of differential misclassification. As a closing comment, one should note that the low κ-value for the comparison of medical and pharmacy records.

The final paper in this section involved a validity-reliability study by Meltzer and Hochstein (18) on those receiving medical care through a health plan. A sample of 1,530 adults answered a survey twice to form the basis of a reliability analysis. The interim period chosen was one week, a compromise between a lengthy period during which there existed the risk of change in the respondents' health and a short period

allowing the respondents to remember their original answers. As in a group of papers discussed earlier, (4,5,6) the object of the analysis was chronic conditions. The difference here was that all questions were yes-no allowing for the use of κ-statistic to correct for change agreement. Chronic conditions were classified as chronic illnesses, disabilities, impariments and symptoms.

Raw percent reliability for the 50,186 total responses was high (96%), but since greater than 90% of the original answers were "no" this meant that 5 out of 6 answers would have been expected to have agreed by chance alone (determined from marginal totals of 2 x 2 contingency table). The raw value, corrected for chance agreement, was .82. Chronic illnesses (κ_{rel} = .89) were more reliably reported than were disabilities (κ_{rel} = .80), impairments (κ_{rel} = .82), or symptoms (κ_{rel} = .79). The reason could be that the subjects learned diagnostic labels of chronic diseases from their physicians, while their responses to the more subjective questions on other types of complaints tended to be reflective of their day to day sense of health.

The validity analysis compared survey responses to medical records. The shortcomings of such a validity base were addressed in the other chronic condition studies (4,5,6). To reduce subjectivity and ambiguity in this case, several steps were taken. First, chronic conditions were subdivided into illnesses, disabilities, impariments, and symptoms - as described in the reliability analysis. This was similar to the Class I - III scheme used by the USNHS (6). The "illness" subdivision here was equivalent to Class I checklist conditions of the earlier study. Secondly, only those subjects with multiphasic (complete) examinations were included. Finally, two abstractors (independent

physicians) were used to summarize all the medical records. Specifically, the abstractors answered the same 40 questions the respondents were asked, using only information contained in the records. No other physicians or questionnaire answers were consulted.

A total of 729 record abstractions were compared with the initial questionnaires using the κ-statistic. The raw agreement figure was 92%. Correcting for chance agreement, the figure plunged to .37. κ_{val} for illnesses was .52; for disabilities, .45; for impairments, .31; and for symptoms, .28. The higher figure for illnesses was due to the same factors that caused Class I checklist conditions to be better reported in the USNHS study (6) - seriousness, specificity, etc. In terms of the earlier chronic condition studies, 48% of all chronic conditions abstracted from records were reported in the surveys, while 54% reported in the surveys were in the abstracted records. Illnesses tended to be underreported with other conditions tending to be overreported.

The overall conclusion, as in the earlier studies (4,5,6), was that the survey information did not agree closely with medical records. Also, as earlier, there was little relation to demographic factors. The results in this case though were obtained through more stringent procedures and thus could be trusted to a greater degree. Curiously, the reliability was considerably higher than the validity in this study. Intuitively, one would not be surprised if an unreliable self-reporter was invalid, but a reliable reporter would not necessarily, as seen here, have to be valid.

MULTIFACTOR VALIDITY (OR RELIABILITY) - INDIVIDUALLY BASED MEASUREMENT

The single paper in this section by Pecoraro et al. (19) resem-

bled those discussed recently in that it continued to employ the κ-statistic methodology in the determination of validity and reliability. There was, however, one major departure from not only these recently analyzed papers but from all those discussed up to this point. The difference was that 23 patients (22 male) were tested individually for reliability and validity in filling out self-administered health history questionnaires. The procedure required no interviewers or preparation on the part of the patients. The reliability analysis consisted of a comparison of the same questionnaire completed by the same individuals two days apart. Note the danger of original answer influence. Validity was defined here as a comparison of either questionnaire (first or second) with a subsequent verbal history, covering the same areas, taken by a physician. The shortcomings of this type of validity base were noted by the authors when they contrasted it with so-called unachieveable "ultimate validation." However, in the real world, validity is a relative concept and could very well have been achieved functionally, if indeed it wasn't achieved with their method, provided the common pitfalls, mentioned on numerous occasions, were avoided. At any rate, the questions were objective, allowing for the simple κ-statistic analysis. The sample of 23 was found to be representative of the particular hospital population from which it was drawn in terms of diagnostic case mix, mean age (58.8 years, range: 37 to 80 years), and sex. As in most cases, the individuals were volunteers and therefore more willing than the average individual to comply. The questionnaire required 115 responses (117 for the woman). The participants knew they would be tested twice but were not informed about the physician interview. Agreement for reliability and validity was defined to be the proportion of consis-

tent answer pairs corrected for chance agreement.

Mean raw individual reliability was 89.6% (range : 75 to 96%) while κ_{rel} was 0.705 (range : 0.495 to 0.883). Mean raw individual validity (first questionnaire) was 92.9% (range : 85 to 98%) while κ_{val} was 0.794 (range: 0.371 to 0.965). No demographic breakdown effects were mentioned. The small sample size was an obvious detrement. In this case note that the mean validity and reliability scores were at the same general level. Note also how the chance-corrected scores differed from the raw values. However, the κ-value for validity was still relatively high - nearly 80%. Most important, though, was the fact that accuracy was defined according to the individual.

OVERVIEW OF LITERATURE AND OBJECTIVES OF THE PRESENT STUDY

The literature reviewed here concerned the most recent and pertinent investigations on the accuracy of self-reported information, the major data source in retrospective epidemiologic studies. A tripartite hierarchical system was employed in presenting the papers. One hierarchy was based upon the complexity of the self-reported information to be collected. Three levels operated here: 1) the number of items to be reported, 2) single vs. multidimensional nature of such items (e.g. simple occurrence vs. occurrence plus duration), and 3) time lapse, i.e. present vs. past occurrence of item(s). A second approach was based upon the methodological techniques that have been used to analyze validity. The progression moved from the use of crude tools such as simple percentages and raw matching to relatively more refined statistical techniques such as correlation coefficients and the κ-statistic. A third was the progression from the group to the individual. Along the

way the methodological and executional strengths and weaknesses of the works discussed were evaluated. The above was based upon what the studies were attempting to accomplish. Quality of data, especially validity bases, was analyzed. Possible restrictions and conversely, general applications, were pointed out. Also considered were important factors affecting the validity and reliability determinations.

These papers, for the most part, represented the available literature on the validity of self-reporting epidemiologic studies, specifically those requiring reporting of multiple or multidimensional items. The lack of literature on this subject was noted by Schlesselman (20) in his volume on case-control studies. What is available also falls short in two respects. First of all, it overwhelmingly opts for using the group method (collective scores) rather than taking on the more difficult but potentially superior and more versatile individual approach (personal scores). Secondly, it is limited to those situations fitting into the aforementioned hierarchy. A step above this hierarchy would be a situation involving the accuracy of self-reported, detailed occupational histories. A fourth level of complexity of collected information could exist in this case - order of a string of events. This, combined with the individual approach, would require a new, possibly difficult, and highly complex methodology of validity determination. Not withstanding the difficulties, there would be a practical reason for developing such a methodology - Detailed occupational histories are commonly used in obtaining exposure information in retrospective epidemiological studies.

With the above in mind, a study was designed to assess the accuracy of ordered and detailed, self-reported exposure information in the

context of an actual retrospective epidemiologic study. Specifically investigated were the individual validities of self-reported occupational histories of PCB-exposed workers. This project was derived from a larger retrospective survey study of PCB effects and occupational exposure by Fischbein et al. (21). The preliminary objectives were to : 1) define validity operationally, 2) based upon this definition design a quantitative measure (score) for it, and 3) obtain these scores for each individual in the study group. There were three major objectives. The first was to see how the scores distributed and if inaccuracies, as measured by lack of validity, lead to misclassifications in the actual PCB survey study (local effects). The second involved statistical manipulations, followed by analysis, of the local data, adjusting for various independent factor effects (possibly confounding) on accuracy to determine the conceptual value of the self-reported work history in assessing exposure status in the generalized self-reporting situation. The third and final objective was to discuss the possible ability of reliability to predict accuracy. This was the important factor in terms of practical study-specific self-reporting accuracy testing.

CHAPTER II: MATERIALS AND METHODS

This particular study had its origins in a repeated survey study (Fig. 1) by Fischbein et al. (21) to assess both the prevalence of health effects associated with occupational PCB-exposure and health status changes resulting from modifications in exposure conditions.

In their study, the target population was the approximately 800 individuals employed (as of March 1976) in labor-intensive operations at two U.S. capacitor-manufacturing plants where PCBs were used. Typically, this group consisted chiefly of blue-collar workers. Also included, however, were foremen and some directly involved craftsmen, engineers, and lower management personnel. From this target population they selected a sample of 326 individuals to participate in an initial PCB-effect prevalence survey (March 1976). An original criterion for selection was ten or more years of employment, i.e. to insure sufficient time for significant PCB-exposure. This, however, did not yield the desired sample size and was modified to produce the above figure. All labor-intensive departments were represented in this survey so as to include individuals with varying intensities of PCB-exposure. The sex breakdown was 168 males and 158 females.

To assess PCB-exposure, each individual was interviewed concerning his or her work history at either of the two plants and had a serum sample taken. The specific period of interest in the work history was from 1947 to 1976 when PCBs were used in the manufacturing process. Up until 1971, the higher molecular weight PCB mixtures, Aroclors 1242 and 1254, were used (HPCB period). The lower molecular weight mixtures, Aroclors 1221 and 1016, replaced them for the remaining five years (LPCB

period). Possible toxicity differences between HPCBs and LPCBs were the reason for making this distinction. The work history information was combined with average departmental PCB air level measurements to produce a work-specific exposure figure. The serum samples were analyzed for HPCBs and LPCBs by gas chromatography to yield a serum-based exposure figure.

Questionnaires concerning past medical history, personal habits (i.e. dietary, smoking, and alcohol consumption, etc.), family history, and reproductive history were administered to identify factors that could have possibly confounded any association between health effects and PCB exposure levels.

In terms of measuring outcome (PCB health effects), each individual had a complete physical examination, completed a questionnaire which stressed symptoms related to pertinent organ systems (with the aid of a physician), and underwent a broad spectrum of laboratory tests.

Considering all the health effect possibilities, there were few findings of correlation between such effects and PCB-exposure levels. One such finding was a significant association between abnormal dermatologic findings - a typical aromatic chlorinated hydrocarbon induced health effect - and serum HPCB levels (Chi-square, $p < .05$, $df = 4$). PCBs have also been known to produce liver abnormalities, but only one indicator of liver function, SGOT, had abnormal levels significantly associated with increased serum PCB levels (HPCBs:Chi-square, $p < .01$, $df = 1$). No work-specific exposure category correlations with health effects were reported although these categories did possess significantly different mean plasma PCB levels.

A followup survey was conducted in December 1979. Nearly 60% of

the 1976 group was reexamined. Also included were 58 new individuals, some of whom started at the plants after PCBs were discontinued. The procedures for obtaining exposure and outcome information were the same as in the initial survey. This time, in addition to the prevalence of health effects, health status changes (from 1976 to 1979) were investigated. As of this writing no results of correlative analyses are available.

The repeated surveys of Fischbein et al. (21) supplied two items that represented the critical body of information for the present study - the population and accompanying work history data. This new study, though, unlike its progenitor, was only concerned with the exposure data - specifically, the accuracy of PCB exposure information as provided by the work histories. Thus, the outcome and PCB serum level data of the original study were relegated to marginal importance. Accuracy determination was made possible by the subsequent (1980) acquisition of a validity base (company employment records) for the members of the survey groups. These records are described in detail under the SOURCES OF DATA subheading which follows shortly.

As described earlier, in the introduction section, the immediate emphasis in this study was on the degree of inaccuracy in self-reporting in the survey groups and if this was reflected in misclassifications in individual PCB-exposure profiles. In terms of the Fischbein et al. survey groups, if the misclassification rates were severe, any of the outcome-exposure associations they found earlier (see p. 39) or those which may be uncovered by them in the future, would be open to question. The ultimate aim, though, in the design of this study, was through statistical manipulations of the local data, to provide insight into

generalized accuracy in retrospective self-reporting studies.

THE STUDY POPULATION

The ultimate makeup of the original main 1976 validity-evaluation group resulted from two processes of selection. The first of these was carried out by Fischbein et al. (21) in their initial PCB-effect prevalence survey where they sampled the 326 individuals from the target population (workforce of 800). To reiterate, the major portion of this group was made up of blue-collar workers. The others were foremen and directly involved white-collar personnel. The second selection occurred under the auspices of this study and was dictated by the availability and legibility of company employment records, the base of validity. Of the original sample of 326, 288 (145 males, 143 females) met the second selection criteria and were thus eligible for the main 1976 validity analysis. The loss of the 38 individuals was due mainly to the unavailability of the records. In several cases, the records were available but illegible. There was a systematic process operating in the case of the lost males. The entire complement of foremen and white-collar personnel (17 individuals) was included in this group of 23. The loss of the 15 females, on the other hand, appeared to be random in that all the females were blue-collar workers. The net result was a more or less homogenous blue-collar group with equal numbers for each sex.

The subset of the original 1976 main validity group which was used in this study to evaluate the effect on validity of examinational delay had to meet one additional criterion - reexamination. The fact of reexamination was established in the December 1979 followup survey by Fischbein et al. (unpub.). The total number of individuals with both

1976 and 1979 examinations plus available company records was 165 (80 males, 85 females), 57.3% of the original 288. Any differences, including their possible effects, between the main and subgroups will be dealt with later, in the analysis chapters.

SOURCES OF DATA

Occupational histories - The function of the occupational history for the surveys of Fischbein et al. (21) was to provide an estimate, hopefully correct, of PCB exposure for each individual. Each individual's occupational history basically consisted of a dated chronological listing of all recalled PCB-period job categories (Figs. 2 and 3). The dates, representing initiation of PCB-period employment, categorical changes, and end of PCB-period (March 1976), were usually specific to the month. In cases where only a year date appeared (rare in initial survey, more common in followup) the entry was designated (for this study) as "6/yr."

Many, but not all the occupational histories, contained for each job category an operational description of varying detail, including supposed exposures (PCBs and others - e.g. TCE, lead, asbestos, etc.) and protective measures taken (e.g. respirators, gloves, goggles, etc.). Such descriptions were at times useful in determining the intended job category since the category alone tended to be ambiguous. Often not included in the occupational history were those intervals during which the individual was off the job, i.e. missed regular worktime. On the reverse side of the history form was a less detailed previous work history.

All information was obtained through the aid of personal inter-

viewers. There were nine employed for the initial survey and seven for the followup. Interviewers had the potential to influence the abovementioned details provided by the individual (interaction between individuals and interviewers). The two interviewer groups were mutually exclusive with the exception of one interviewer who worked both surveys.

In all, there were approximately 40 job categories, including "not working" (Table 1). However, most of the listed jobs fell into 20 of these (main operation jobs plus general maintenance). Note the column on the extreme right of Table 1. This consisted of the inherent PCB risk factors for each job category based upon an industrial hygiene survey by Jones et al. (22). The inherent risk factor for a particular category multiplied by the number of workday months (21 to 22 days per workday month) spent in that category represented an individual's total risk for that category. Summing up the total risk figures for each category yielded an individual's self-reported overall or cumulative total risk. Separate overall total risks were obtained for each PCB-mixture period (Fig. 5: 1947-71 for HPCBs, 1971-76 for LPCBs). Overall total risk figures for each individual were then assigned to one of four exposure categories (based upon quartiles: very low, low, medium, and high).

Personnel records - The personnel records (company employment records), which made accuracy determination possible, were contained on a series of twenty microfilm cartridges. The sorting was alphabetical, by surname. They contained complete work records at the two plants for each individual (Fig. 4). Unlike the occupational histories, these records provided a chronological listing of not only jobs to which the worker was assigned, but time lost to illness, layoffs, strikes, disciplinary actions, military service, or maternity leave, etc. All infor-

mation was accurate to the day permitting exact determinations of job starting and ending dates and duration. Each job listing had a short description and four-digit code from a master list containing all company jobs. This code had to be translated into the categorical system used by Fischbein et al. in their survey study. Also included in the records was a notation of the appropriate departmental foreman - a valuable cross-reference in case of ambiguous listings.

The information contained in these records was obtained at the time of occurrence and from a source independent of the workers themselves (did not depend upon recall ability, cooperativity, truthfulness, etc., nor upon interviewers) and thus represented a relatively objective base of validity (actual value, reference standard). There was no evidence to suspect that these records were altered or biased since their function seemed to be more geared toward bookkeeping (keeping track of wage categories for financial, disciplinary, promotory, or legal purposes) rather than for recording PCB-exposure. Any bookkeeping flaws that might have occurred could not be checked.

In terms of PCB exposure, the same procedure was used with these records as with the occupational histories to obtain overall total risk figures and risk categories for each individual. These record-based figures and categories of course represented the actual (relatively speaking) situation unlike those derived from the work histories which were subject to error. Additional information provided by these records was sex, date of birth, height, weight, marital status, educational level, other family members, and previous employment.

It should be noted that these records had one shortcoming. While being an accurate source of information concerning the assigned job ca-

tegory or department, i.e. where the worker should have been, they might have missed variations within the defined categories (in terms of PCB-exposure), unauthorized job shifts (e.g. trading between workers), and overtime, which could have been picked up by the occupational histories. Notwithstanding the above, the choice of company records as the reference standard was made on a relative but practical basis, considering real world conditions.

JOB CATEGORIES

The blue-collar operation at the two plants consisted of nearly 300 job titles (master list). Each one of these had a four-digit code. The first two digits of this code represented either a department or a major skill type, while the last two indicated internal subclasses. For this study (and the original work of Fischbein et al.) a new list of job categories was created (Table 1). The production process was split into three major divisions: the main operation (PCB-capacitor production), ancilliary operations, and plant maintenance. Within each of these divisions were specific job categories. They were defined with an emphasis on the production process (three-digit code) with letter subscripts designating particular individual skills that segregated the workers into a specific recognizable class within a production area. The main operation had ten major categories with 18 total classes. Note that "not working" was also treated as a category. The purpose of this new type of classification was to facilitate measurements of PCB risk factors (Table 1, column 4). These risk factors were more readily determined on a job categorical basis rather than on the individual scheme of the master list. An additional advantage of this condensed list was

the fact that it made the analytical process more manageable.

OPERATIONAL DEFINITION

Validity (crude score) is defined as the percent agreement between either the 1976 or 1979 occupational history and the company employment record for a particular individual (Eq. 1 and 2).

DETERMINATION OF CRUDE SCORES

The process began (after translating company job codes to the system in Table 1) with the listing of the earliest date of employment (6/47 or later), whether this appeared on the company record or either occupational history (Fig. 5). In Figs. 2, 3, and 4, this date was the same, 1/68. This date represented the beginning of the first comparison interval. This interval ended with the earliest job-change date, once again, whether this occurred in Figs. 2, 3, or 4. This date, in each figure, was 3/68. Interval "1" was thus from 1/68 to 3/68. For these two months the job category according to each source (company record, 1976 history, 1979 history) was recorded. Either history could agree or disagree with the company record listing of 007C (washing) for this interval. Both histories in this case agreed with the company record. The next job-change date occurred in the company record (Fig. 4) at 6/68. Interval "2" thus spanned the three months between 3/68 and 6/68. For this interval, once again, there was agreement for each comparison (007A: treat). The first disagreement was in interval "4" for the 1979 history. The process continued through 16 intervals to 3/76, the PCB discontinuation date. The total number of months in the recall period (1/68 to 3/76) was 98.

Overall percent agreement = total # months contained in agreement intervals / # months in recall period

Eq. 1: VAL_{76} = total # months contained in agreement intervals [(Job(76) = Job(CRt)] / # months in recall period

= 57/98

= 58.2

Agreement intervals: "1-4, 6, 13, 14, 16" = 57 months

Eq. 2: VAL_{79} = total # months contained in agreement intervals [Job(79) = Job(CRt)] / # months in recall period

= 56/98

= 57.1

Agreement intervals: "1-3, 7, 8, 14" = 56 months

Crude validity scores were obtained for 288 individuals in 1976, the initial examination period. The followup examination in 1979 included 165 of the original group. The entire crude analysis is summarized in Fig. 6. In addition, the crude scores of those whose work histories went as far back as the 1950s were broken down into three subscores, one for each of three subperiods: 1950-59, 1960-69, 1970-76.

The method of crude score determination just described allowed for two different error types to result in non-agreement. Specifically, the degree of non-validity was a function of both job categorical disagreements in fact (and duration) and in chronology (order, change dates). The question could have been raised as to why a measure was devised that was sensitive to both of the above error types when a self-

reported work history could have been invalid merely on account of incorrect categorical change dates while agreeing with the validity base on the fact of time spent within each job category. The answer is that it was necessary because exposure conditions at the two plants changed during the period of investigation, 1947 to 1976. Changing conditions imply that job categorical risk figures are functions of time and thus the same job category under different conditions of exposure may be a different entity. In essence, a sufficiently large chronological error could become a categorical error. The target population of Fischbein et al. (21) which supplied the study group for this investigation was subject to two distinct subperiods of exposure conditions, 1947-1971 and 1971-1976. These were defined by the switch in PCB mixtures in 1971. Other changes, difficult to pinpoint, must also have occurred. These involved protective measures, government or company exposure allowances, and/or worker awareness of PCB-toxicity, etc. In any retrospective study involving workplace exposure (This study is concerned with the generalized as well as local validity of the work history.), one would be surprised if plant conditions did not vary to some degree over time. Therefore, a validity score sensitive to chronology errors would be appropriate for any retrospective occupational validity study.

MISCLASSIFICATION DETERMINATION

Each occupational history was compared to the validity base (company records) in terms of the PCB-exposure categories described earlier (i.e. quartiles based upon overall total risk). The percentage of individuals for which there was categorical (exposure class) agreement was recorded and defined as validity of the exposure category. Disagree-

ments were recorded as either upward or downward shifts (relative to the validity base) and represented misclassifications. Separate measurements were made for 1976 and 1979 and the HPCB and LPCB periods.

FORMAT

All calculated validity scores, overall total risk figures, risk categories, and personal interviewers, along with various individual demographic and occupational variables, were entered into a computer file, using the CUNY (WYLBUR) editing system. To facilitate analysis, only numeric characters were used. The above information for each individual spanned three 80-column cards. The list of variables with column numbers and appropriate explanations is given in Table 2.

ANALYSIS

SAS procedures were used for all statistical analyses reported in this study. The analysis is divided into three chapters. The first of these involves a presentation and discussion of descriptive findings concerning specific personal demographic and occupational variables. The continuous variables were separated in terms of appropriate class variables and presented in terms of frequency distributions, mean, standard deviation, and standard error of the mean. Student's t-test (paired or unpaired, when appropriate) was used to evaluate the significance of any observed class differences in the continuous variables.

The next chapter is concerned with the presentation of both local findings and generalized implications concerning the accuracy (validity) of retrospective self-reported information. The local findings were all

descriptive in nature and any observed class differences were evaluated by Student's t-test, ANOVA, McNemar's test, or the sign test, when appriate. The procedure for the generalized validity analysis, where the effects of measureable independent variables (including the above personal variables) were considered, was multiple linear regression. The particular type of regression analysis employed here made full use of the continuous data and allowed examination of a response affected by both qualitative and quantitative independent variables. The quantitative variables (possible confounders) were adjusted for so that in addition to their own effects on the mean response (validity) the effects for each qualitative (class) variable could be compared in an unbiased manner. It also permitted the similar examination of possible interactions between the qualitative and quantitative variables.

The final chapter contains the discussion of all findings from the previous two chapters.

Table 1. Job categories at the two capacitor-manufacturing plants

Job Category	Code	Description of Operation	Inherent PCB Risk Factor
Not Working	000	Layoff, illness, accident, strike, pregnancy, suspension, leave of absence, military service, etc.	None
Unknown	001	Unclear or missing job description	
		MAIN OPERATION	
Foil Mill	002	Roll the foil used in the core of the capacitor	None
Casemaking	003	Two-inch slugs of aluminum are heated, melted, and formed into cases; large storage bins of cases are often shaken, generating large amounts of dust; cases are degreased in TCE	.035
Winding, Tap	004	Paper and foil are wound together to make the core of the capacitor; enclosed room	.082
Cover	005	Assemble covers for capacitors; enclosed room	.007
Assembly (HF) Assembly (FE)	006A 006B	Rolled core is placed in the case: press the roll, weld and solder taps and cables to the cover; seam unit to enclose top and bottom; fill hole is left open in the top; PCBs are used to lubricate (applied by brush or fingers 5 to 6 times per day -- a 1 to 2 minute operation)	.052 .440
Epoxy	006C	Tubular assembly	.052

Table 1. Continued

Job Category	Code	Description of Operation	Inherent PCB Risk Factor
Treat	007A	Capacitors are filled with PCBs and heat treated; men read gauges to control temperature and the rate at which the capacitors travel through this enclosed process; PCBs are hot and fuming; spills occurred in the past; degreasing takes place here	.937
Sealing	007B	Fill holes are sealed	.600
Washing	007C	PCB-covered units are washed in water and then stacked	11.000
Testing	008A	Capacitors are loaded onto conveyor system which moves them to the testing machine; purpose is to test for leaks	.410
Painting	008B	Spray-painting of capacitors	.260
Label and Pack	008C	Stamp, label, and pack capacitors for shipping	.260
Rework	009A	Salvage operation to repair defective units and remove usable parts	.410
QC (FE)	009B	Quality control	.410
QC (HF)	009C		None
Repair Cap.	010	Old repair operation in treat area	.600
Shipping	011		None
		OTHER AREAS	
Film	012	Dielectric process operators	None
Battery	013	Assemble battery cells	None
Boxmaking	014		None

Table 1. Continued

Job Category	Code	Description of Operation	Inherent PCB Risk Factor
Typograph	015		None
FKC Switch	016		None
Insulation	017		None
Pilot Plant	018	Developmental laboratory	None
Microminiatures	019	Old operation (relocated in 1967)	None
Clerical	020	Office work	None
Dispatching	021		None
Machinist	022	In enclosed shop	None
		MAINTENANCE	
PEC	023	Plant engineer craftsmen	Var.[a]
Helper	024	Maintenance aide	Var.
Security	025		Var.
Electrician	026		Var.
Welder	027	Maintenance welding	Var.
Painter	028	Maintenance painting	Var.
Elevatorman	029		Var.
Janitor	030		Var.
Oiler	031	Lubrication maintenance	Var.
Boilerman	032	Boiler maintenance	Var.
Trans. Equip.	033	Maintenance of transportation equipment	Var.

[a]Depends upon area(s) in which work occurred

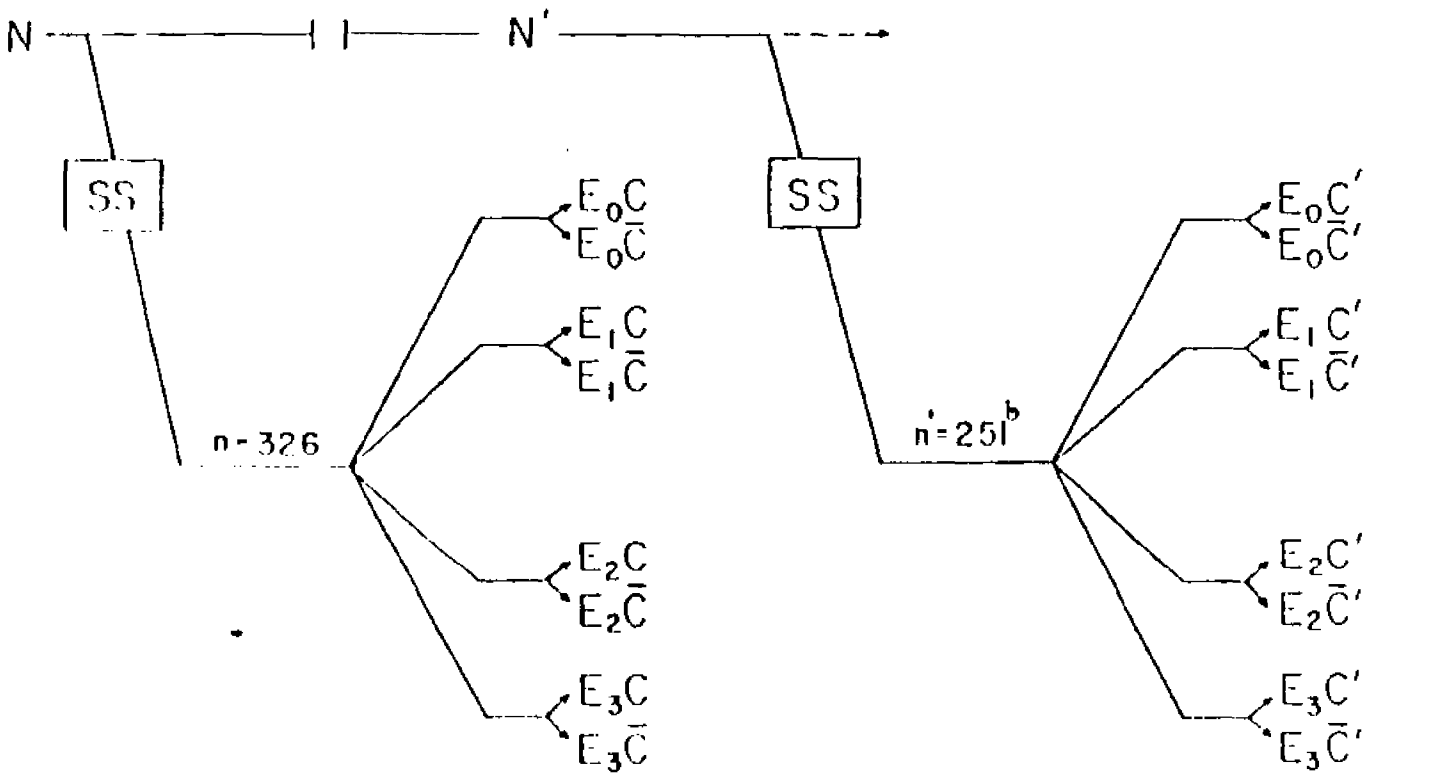
Table 2. List of variables, column numbers, and brief explanations

Variable	Column#	Explanation
ID Number	1-4	Four-digit identification number for each individual
Age (as of 3/76)	6-7	
Sex	9	1=male, 2=female
Educational Level	11-12	Highest grade completed: 1-16
Marital Status	14	1=single, 2=married, 3=divorced, 4=widowed, 5=remarried
1976 Interviewer	16-17	Nine different interviewers: 1-9
1979 Interviewer	19-20	Seven different interviewers: 1-7
Recall Period	22-24	Period to be recalled by subjects for crude validity determination
Duration (CR)	26-28	Total months actually worked during the PCB period (according to the validity base)
Duration (76)	30-32	Total months worked during the PCB period (according to the 1976 work history)
Duration (79)	34-36	Total months worked during the PCB period (according to the 1979 work history)
Validity (76)	37-40	Crude 1976 validity score
Validity (79)	41-44	Crude 1979 validity score
Validity (76): 3	53-56	1950-59 subscore
Validity (76): 2	57-60	1960-69 subscore
Validity (76): 1	61-64	1970-76 subscore
Job Diversity Index	106-108	Complexity of job categorical pattern
Examined in 1976	110	1=yes, 2=no
Examined in 1979	112	1=yes, 2=no

Table 2. Continued

Variable	Column#	Explanation
Cum. Exp. (71): CR	114-117	Cumulative exposure up to 1971, HPCBs, according to validity base
Exp. Cat. (71): CR	119	Exposure category up to 1971, HPCBs, according to validity base
Cum. Exp. (76): CR	121-124	Cumulative exposure from 1971 to 1976, LPCBs, according to validity base
Exp. Cat. (76): CR	126	Exposure category from 1971 to 1976, LPCBs, according to validity base
Cum. Exp. (71): 76	135-138	Cumulative exposure up to 1971, HPCBs, according to 1976 work history
Exp. Cat. (71): 76	140	Exposure category up to 1971, HPCBs, according to 1976 work history
Cum. Exp. (76): 76	142-145	Cumulative exposure from 1971 to 1976, LPCBs, according to 1976 work history
Exp. Cat. (76): 76	147	Exposure category from 1971 to 1976, LPCBs, according to 1976 work history
Cum. Exp. (71): 79	166-169	Cumulative exposure up to 1971, HPCBs, according to 1979 work history
Exp. Cat. (71): 79	171	Exposure category up to 1971, HPCBs, according to 1979 work history
Cum. Exp. (76): 79	173-176	Cumulative exposure from 1971 to 1976, LPCBs, according to 1979 work history
Exp. Cat. (76): 79	178	Exposure category from 1971 to 1976, LPCBs, according to 1979 work history

Figure 1. Repeat survey format[a]



N = Workforce in March 1976

SS = Selected sample

E_k = Exposure category (k = 0,1,2,or3)

C = Existence of abnormality (in 1976)

$\bar{C}$ = No abnormality (in 1976)

N' = Workforce in December 1979

C' = Existence of abnormality (in 1979)

$\bar{C}'$ = No abnormality (in 1979)

[a]Adapted from Kleinbaum et al. (23)

[b]n' = 193 from sample n plus 58 new individuals

Figure 2. 1976 interview for work history

Name_____ ID Number 1892

OCCUPATIONAL HISTORY -- PCBs

Period	Job Title	Job Description by Task and Location	Exposures	Protective measures
1/68-3/68	Washing	Separate and wash units covered with oil	Pyranol, TCE, Salvatone	5a,9,10a, 11a,14
3/68-4/73	Treat	Sending on units covered with oil	Pyranol	Same as above
4/73-10/75	Stock-room	Storage and withdrawal of cardboard boxes		
10/75-3/76	Foil Mill	Handled foil covered with oil (kerosene)	Kerosene	5,6,9,10, 11,14

Figure 3. 1979 interview for work history

Name_____ ID Number 1892

OCCUPATIONAL HISTORY -- PCBs

Period	Job Title	Job Description by Task and Location	Exposures	Protective measures
1/68-3/68	Washing	FE -- Wash units with PCBs	PCBs	Gloves
3/68-6[a]/69	Treat	Checking units covered with PCBs	PCBs	Gloves
6[a]/69-6[a]/71	Packing	FE -- Shipping area: accumulate, sort, pack, and ship	Nothing	
6[a]/71-6[a]/73	Assembly	FE -- Assembled capacitors	TCE	
6[a]/73-3/76	Stock-keeper	HF -- Storeroom: handled cardboard	Nothing	

[a]Only year date given

Figure 4. Company employment record (validity base)

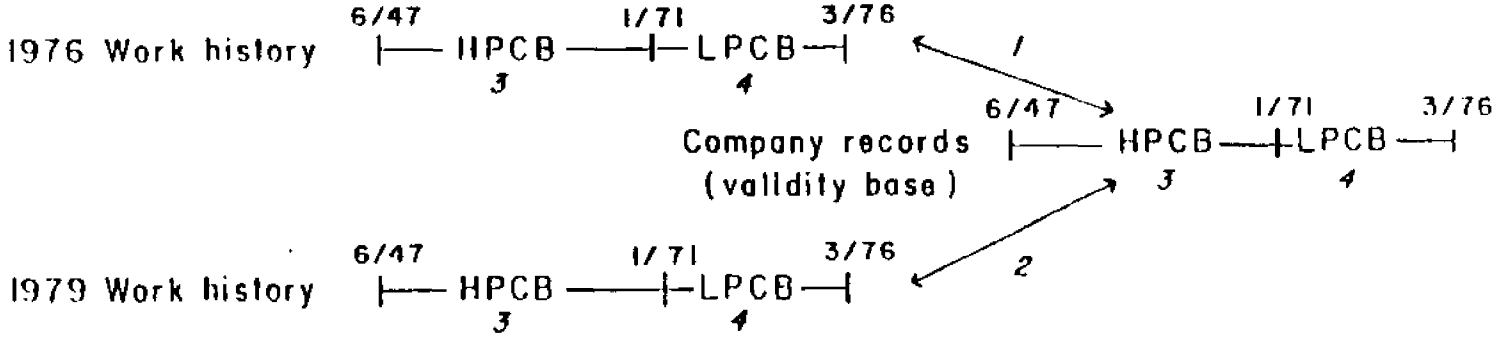
Date	Eng.-Trans. Rem.	Code	Description	Wage Rate	Foreman
1-8-68	Eng.	5405	Det. Wash -- load and maintain		
3-11-68	Trans.	1415	Salvage Repair -- Treat area		
6-13-68	Trans.	6205	Check and Count Heavy -- Treat area		
10-10-69	Rem.	0000	Strike		
2-4-70	Reeng.	6205	Check and Count Heavy -- Treat area		
4-2-70	Trans.	0933	Material Handler -- Stockroom		
1-15-70	Trans.	0932	Transtacker Operator -- Stockroom		
4-9-71	Trans.	4631	Setup Class C -- Winding area		
4-14-72	Trans.	1731	Test Press Operator -- Test area		
6-19-72	Trans.	0601	Stockkeeper -- Stockroom		
9-16-75	Trans.	1901	Foil Operator B -- Foil area		

Figure 5. Crude validity score computation worksheet

ID Number: 1892 DUR: 94 months 1976 Interviewer: 6
Age (as of 3/76): 58 JDI: 174 1979 Interviewer: 3
Sex: Male

Intl.	Start	End	Mon.	Job(CR)	Job(CRt)	Job(76)	Job(79)
1	0168	0368	2	5405	007C	007C	007C
2	0368	0668	3	1415	007A	007A	007A
3	0668	0669	12	6205	007A	007A	007A
4	0669	1069	4	6205	007A	007A	011
5	1069	0270	4	0000	000	007A	011
6	0270	0470	2	6205	007A	007A	011
7	0470	1170	7	0933	011	007A	011
8	1170	0471	5	0932	011	007A	011
9	0471	0671	2	4631	004	007A	011
10	0671	0472	10	4631	004	007A	006B
11	0472	0672	2	1731	008A	007A	006B
12	0672	0473	10	0601	011	007A	006B
13	0473	0673	2	0601	011	011	006B
14	0673	0975	27	0601	011	011	011
15	0975	1075	1	1901	002	011	011
16	1075	0376	5	1901	002	002	011
Recall Period			98				
HPCB Overall Total Risk Figure					42.2	53.9	37.1
LPCB Overall Total Risk Figure					4.3	27.0	12.9
Crude Validity Score (%)						58.2	57.1

Figure 6. Validity analysis procedure



VALIDITY ANALYSIS PROCEDURE

1. Validity of 1976 work history
2. Validity of 1979 work history
3. Period of exposure to high MW PCB compounds
4. Period of exposure to low MW PCB compounds

CHAPTER III: ANALYSIS OF DEMOGRAPHIC AND OCCUPATIONAL FACTORS

The following chapter deals descriptively with those measureable independent variables (see Table 2) which possessed the potential to influence or confound this or any generalized assessment of validity. These were the demographic (age, sex) and occupational (duration of employment, job diversity index) attributes which varied varied markedly or differed among the individuals in the study population (one of two criteria necessary for a confounder).

AGE AND SEX

Fig. 7 summarizes the age and sex structure of the original main group. The mean age for the males was 41.7 years. The range in their ages was from 22 to 73 years. Their age frequency distribution, based upon five-year intervals, appeared to be composed of two subdistributions, each skewed to the right. In practical terms there may have been two separate age classes of male workers. The younger group ranged in age from 20 to 39 years and had a modal interval of 25-29 years. The older group included individuals between 40 and 73 years of age and had a modal interval of 45-49 years. The mean age for females was 47.3 years, significantly higher than that for all males (Student's t-test, $p < .0001$). Their age range was from 22 to 69 years. The female distribution of age intervals, unlike that of the males, had a single definite mode at the interval of 50-54 years and was skewed to the left.

Roughly speaking, the losses for each sex from the initial to the followup examination were equivalent. The mean age for males in the reexamined subgroup was 42.3 years. That for females was 46.6 years.

These values represented insignificant deviation from those seen in the original main group. The female mean value remained significantly higher than that for males (Student's t-test, $p < .005$). In terms of the distributional shapes there was also little change. The males retained the same younger and older subdistributions, each skewed to the right and with the same modal intervals. The female distribution was once again skewed to the left with the mode at the 50-54 interval. There was thus no evident systematic elimination of individuals from the 1976 to the 1979 evaluation with regard to age or sex. All of the above information about the reexamined subgroup is summarized in Fig. 8.

DURATION OF EMPLOYMENT

Duration of employment, DUR, represented the actual time spent working (according to company records) at either of the two plants during the PCB period (1947-1976). Excluded from DUR as defined above was active time lost due to illness, layoffs, strikes, pregnancies, suspensions, leaves of absence, military service, and accidents, etc. It was expressed in workday months (i.e. 21-22 days - weekends and holidays excluded).

Table 3 gives the mean, standard deviation, standard error, and skewness values of DUR for the original main group. DUR values ranged between 21 and 352 months. The mean values indicated a longer DUR for the female workers at the two plants. The male mean value was 170.5 while that for females was 179.1. However, the difference between the male and female mean values was not significant. Fig. 9 illustrates the distributional patterns for male and female DUR. The male distribution here again appeared to be bimodal as was the case with age. The

only difference between this pattern and the one seen for age was that analagous long and short DUR subdistributions were more sharply defined and less skewed. The distribution of female DUR intervals was, unlike the distribution of their age intervals, apparently bimodal. A relatively longer DUR subdistribution had a mode in the 230-249 month interval while the one for shorter DUR had its mode in the 100-119 month interval.

The bimodal appearance of the male DUR distribution was no surprise. After all, one would have intuitively expected duration and age to run along similar lines. However, DUR, as defined in this study, represented a net figure - absolute duration corrected for lost working time. No matter, the males here tended to spend most of their potential work time on the job. Lost time was necessarily kept to a minimum by the need to support the family or maintain position or security in the company. Additionally, epidemiologic studies have shown males to have relatively lower morbidity rates than females (24). Whatever time loss did occur was limited and more or less uniform for all males (e.g. strikes). The bimodal pattern for female DUR was surprising. It did not follow the example set by their age distribution. The age-DUR difference, however, was indirectly attributable to the time loss factor. First of all, they had more cause to lose time than did the males and for longer periods. They became ill more often than males and were more subject to layoffs. They were also compelled to take considerable time off for pregnancies, child rearing, and family problems. The net result of the above was a roughly equivalent mean DUR for both sexes in spite of the significant age difference between them. The fact that the female propensity for time loss was not evenly distributed accounted for

the difference between their age and DUR distributions. The workers hired relatively early spent most of the PCB period beyond the ages of childbearing. They contributed more service time and made up the "greater than 220 month" subdistribution. Many of the remaining younger women who belonged to the shorter DUR subdistribution became pregnant and raised children during the PCB-period which explained the attrition in the 140-220 month range.

For the females reexamined in the 1979 followup, the DUR situation was quite static. There was no appreciable change in the mean value. The reexamination mean was 177.4 months, virtually identical to the main group value. The distributional pattern was almost identical to that for the original group with the exception of the shorter DUR subdistributional mode which shifted up one interval to 120-139 months. There was thus no evidence of any selection bias caused by reexamination for this variable in females. For the reexamined males there was some indication of an upward movement in DUR. The mean value increased to 182.3 months in this subgroup. This small upward shift of nearly one year, however, was not statistically significant. The 1976 and 1979 distributions of male DUR intervals did not appear to show any great differences. The same subdistributions with the same modal intervals were present. All of the above information is to be found in Table 3 and Fig. 10.

JOB DIVERSITY INDEX

Intuitively, one would expect that an occupational attribute such as the number of job categorical changes would affect the accuracy of a self-reported occupational history. Logically, an increased number of

such changes would make the actual history more complex and thus regardless of an individual's innate recall ability limit the maximum obtainable degree of accuracy. However, the number of changes alone is not a sufficient determinant of complexity. One must also consider the spatial arrangement of the changes. Two individuals, each with the same number of changes, may have vastly different spatial patterns. One with changes scattered throughout the work tenure, producing many small intervals, would have a more complex pattern than one with one relatively large interval and the remaining changes compressed within a very short time span. Taking both of these factors into account is the Shannon-Wiener diversity index. This index was developed to measure complexity (i.e. diversity = complexity) in ecological communities. In this study, it was designated job diversity index (JDI) and was adapted to measure the complexity of the job categorical patterns in the company employment records. The formula for the JDI computation was as follows:

$$JDI = - \sum_{i=1}^{k} [(n_i/N)\ln(n_i/N)] \times 100$$

where: N = absolute duration (initiation of employment to 3/76) in months

n_i = number of months spent in interval i

k = total number of intervals

A high JDI would be indicative of many job categorical changes and a scattered spatial arrangement, intuitively producing a large negative effect on the degree of accuracy. A low JDI, on the other hand, would reflect fewer job categorical changes and/or one or two dominating

intervals, intuitively interfering less with the degree of accuracy. It is important to note that within the definition of JDI each job category change denoted a new interval whether or not that particular category appeared earlier in the actual work history. The minimum possible value of the JDI would be 0, when k = 1 and thus n_i = N. This would occur if there were only one interval (job category) that lasted for an entire absolute duration. In this study, such a situation never occurred, but one would have expected recall of such to have been virtually absolute. The maximum possible value for the JDI depends upon two factors. It occurs for a given absolute duration when k = N and thus all the n_i values are one month long (i.e. Minimum interval size yields maximum number of possible intervals.). Indirectly, the maximum value also depends upon the size of N. A larger value of N would allow for a greater number of intervals and in practical terms more room for complexity. The absolute maximum did not occur for any individual in this study, but there were some values that approached this level. One would have expected difficulties in reporting accurate histories in such cases. The behavior of the JDI in some selected situations is shown in Fig. 11.

This JDI measure has two possible limitations to consider. First of all, there could be a situation where there are many job categorical changes (causing high JDI), but arranged in a nonrandom pattern (not difficult to recall - e.g. regular change every six months). Secondly, there could be a situation of frequent changes (again causing high JDI), but a pool of relatively few job categories to switch between (decreasing difficulty of recall). Neither of these limitations, though, appeared to be a factor in this study as changes of category appeared to be random and each sex had both roughly equal and a fair number of job

categorical possibilities.

Table 4 gives the mean JDI values for the original main group. Male values ranged from 26 to 288 with female values ranging from 14 to 326. The female mean value of 214 was significantly higher than the male mean of 170 (Student's t-test, $p < .0001$). This indicated that females had considerably more complex actual work history patterns. The distributions were more or less normal (slight skew to the left) except that there was a relative gap in the male 140-179 region (Fig. 12A). Note how much the female distribution was shifted to the right relative to that for males (Fig. 12B).

The JDI came the closest of the three continuous variables discussed here to being normally distributed. This type of distribution was true for both sexes. The similarity, though, ended there as females had significantly higher JDI levels than males and thus more complex actual work history patterns. This higher complexity was definitely influenced by excessive time lost to illness, pregnancies, and layoffs, which intermeshed with what might already have been a more complex categorical pattern caused, for example by females having been intentionally shunted more often than males from category to category - an occurrence not uncommon in the workplace. Their higher frequency of layoffs was evidence of such treatment. No matter what the cause, the sex difference in JDI appeared to be real and may have had important implications.

For the reexamined subgroup there were no significant deviations from the original main group figures in terms of central tendency. The distributions also maintained their original patterns - no systematic selection for JDI by reexamination. Table 4 and Fig. 13 present the JDI results for the reexamined subgroup.

Table 3. Duration of employment in workday months

Original Main Group					
Sex	Number	Mean[a]	Std. Dev.	Std. Err.	Skewness
M	145	170.5	92.1	7.65	0.11
F	143	179.1	84.9	7.10	0.05

Reexamined Subgroup					
Sex	Number	Mean[a]	Std. Dev.	Std. Err.	Skewness
M	80	182.3	90.7	10.14	-0.01
F	85	177.4	84.0	9.11	0.07

[a]Male-female differences were NS (Student's t-test)
Main-subgroup differences were NS (Student's t-test)

Table 4. Job diversity index

Original Main Group					
Sex	Number	Mean[a]	Std. Dev.	Std. Err.	Skewness
M	145	170	51	4.23	-0.24
F	143	214	57	4.77	-0.55

Reexamined Subgroup					
Sex	Number	Mean[a]	Std. Dev.	Std. Err.	Skewness
M	80	170	53	5.93	-0.16
F	85	210	57	6.18	-0.71

[a]Male-female differences for both groups were significant (Student's t-test, $p < .0001$)
Main-subgroup differences were NS (Student's t-test)

Figure 7. Age data for the original main group

A. Males

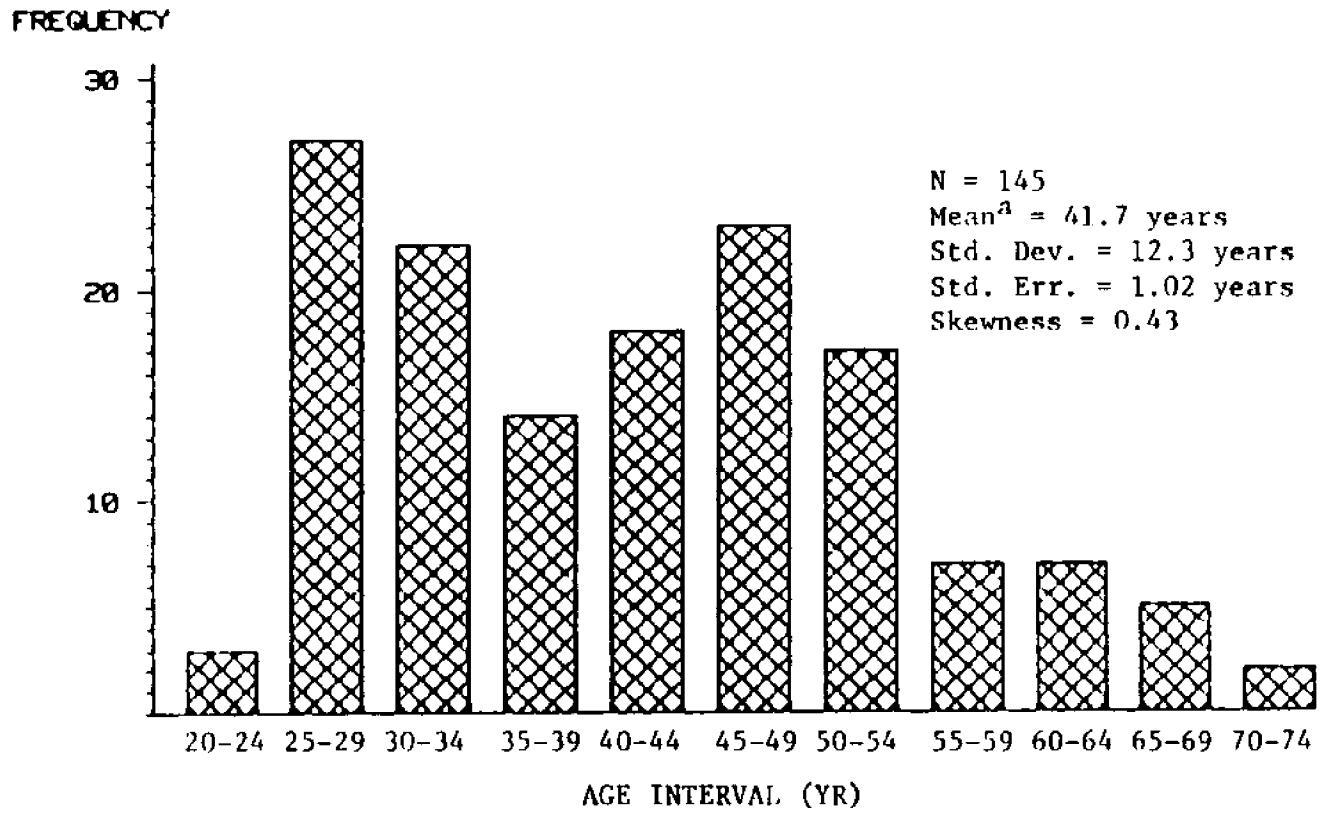
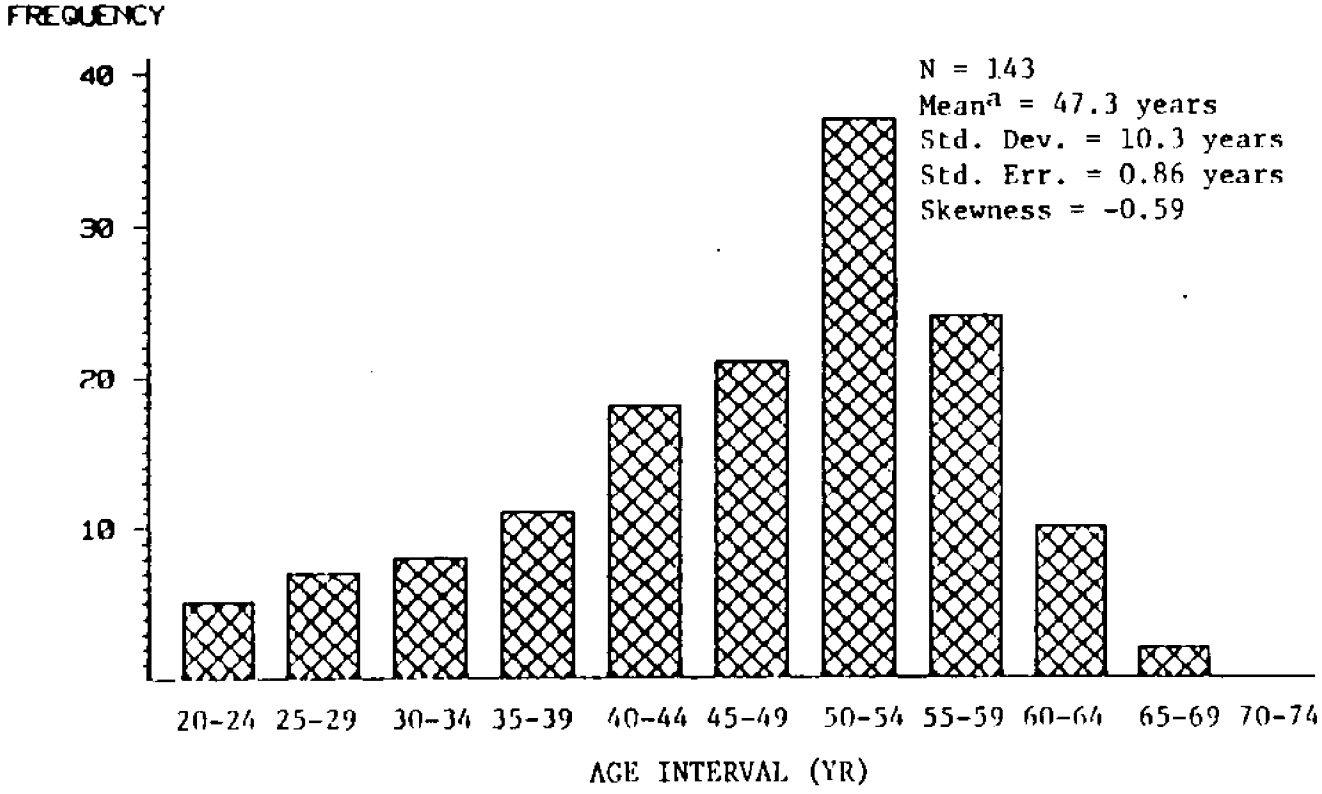


Figure 7. Continued

B. Females



[a]Male-female difference is significant (Student's t-test, $p < .0001$)

Figure 8. Age data for the reexamined subgroup

A. Males

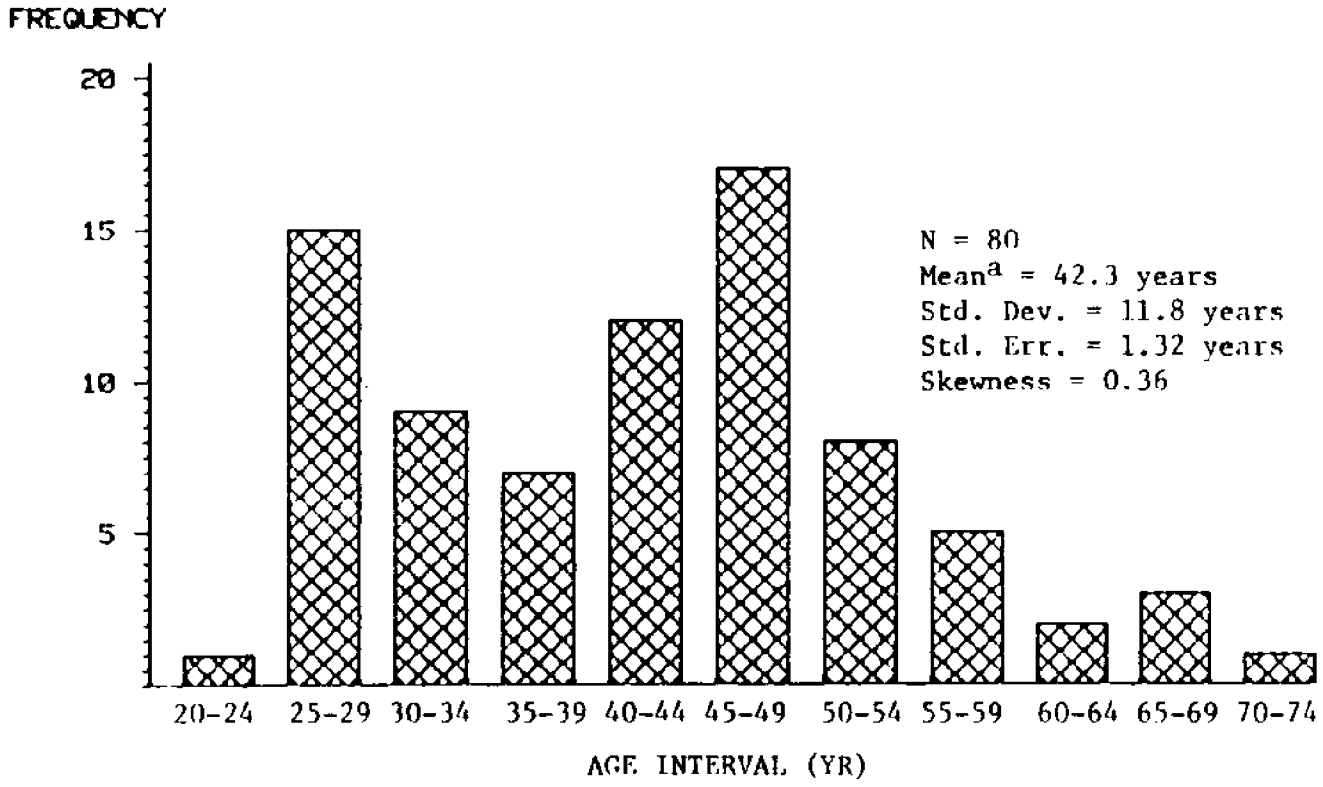
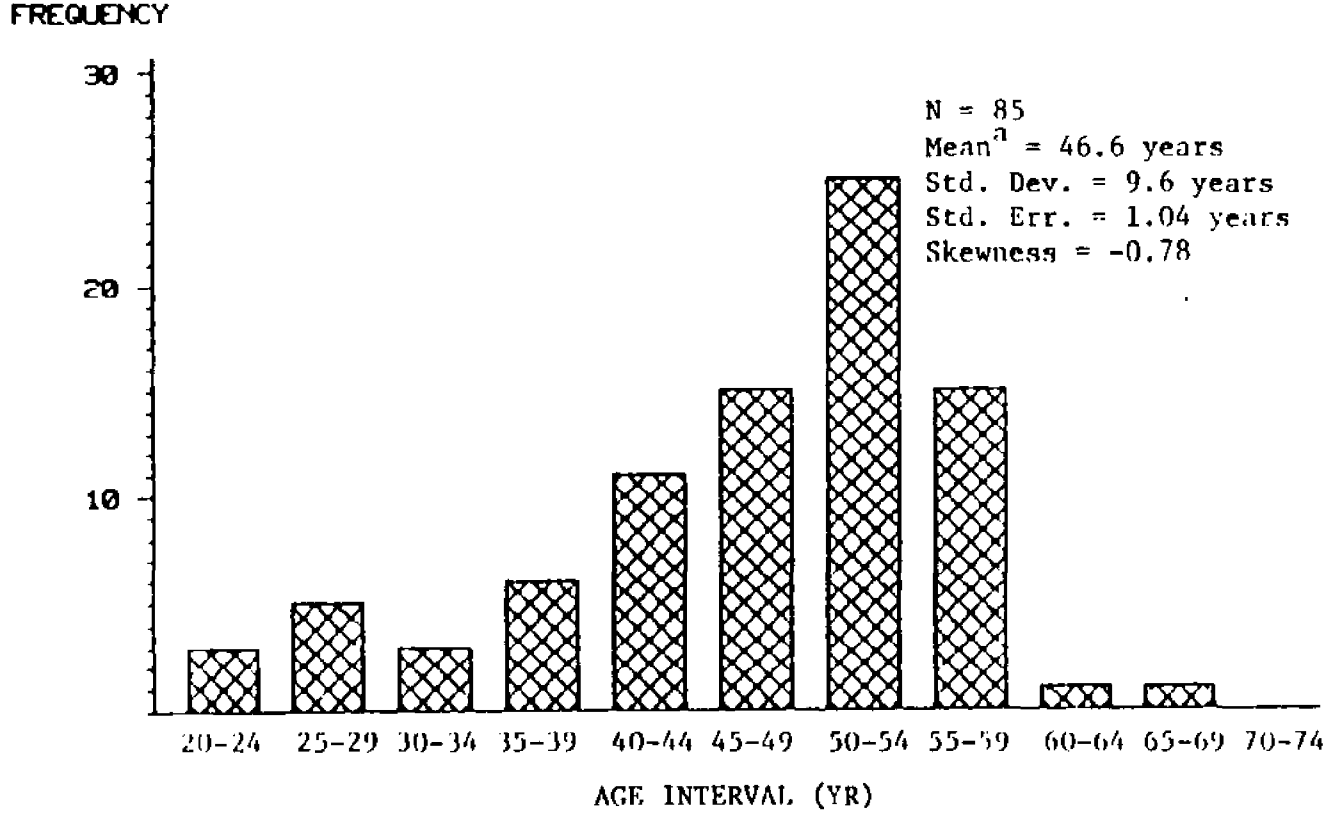


Figure 8. Continued

B. Females



[a]Male-female difference is significant (Student's t-test, $p < .005$)

Figure 9. DUR distributions for the original main group

A. Males

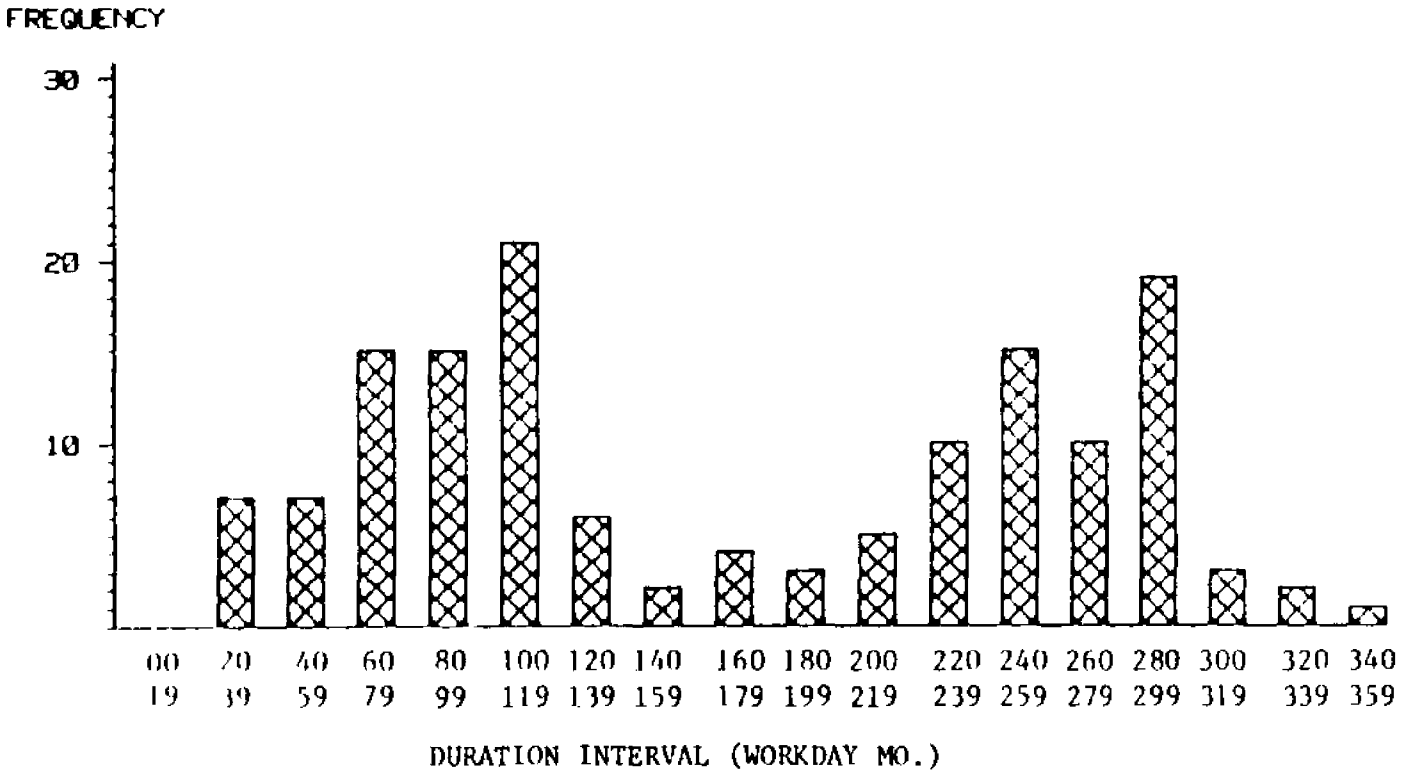


Figure 9. Continued

B. Females

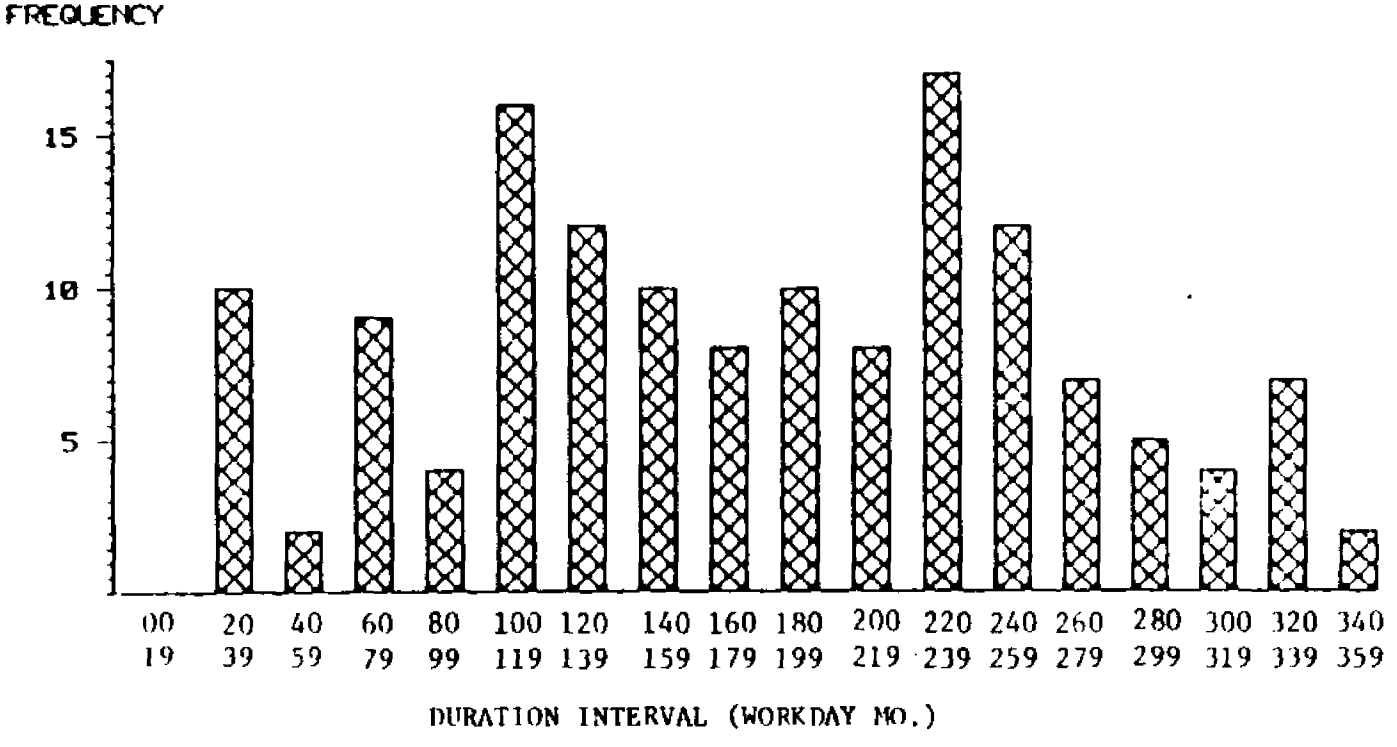


Figure 10. DUR distributions for the reexamined subgroup

A. Males

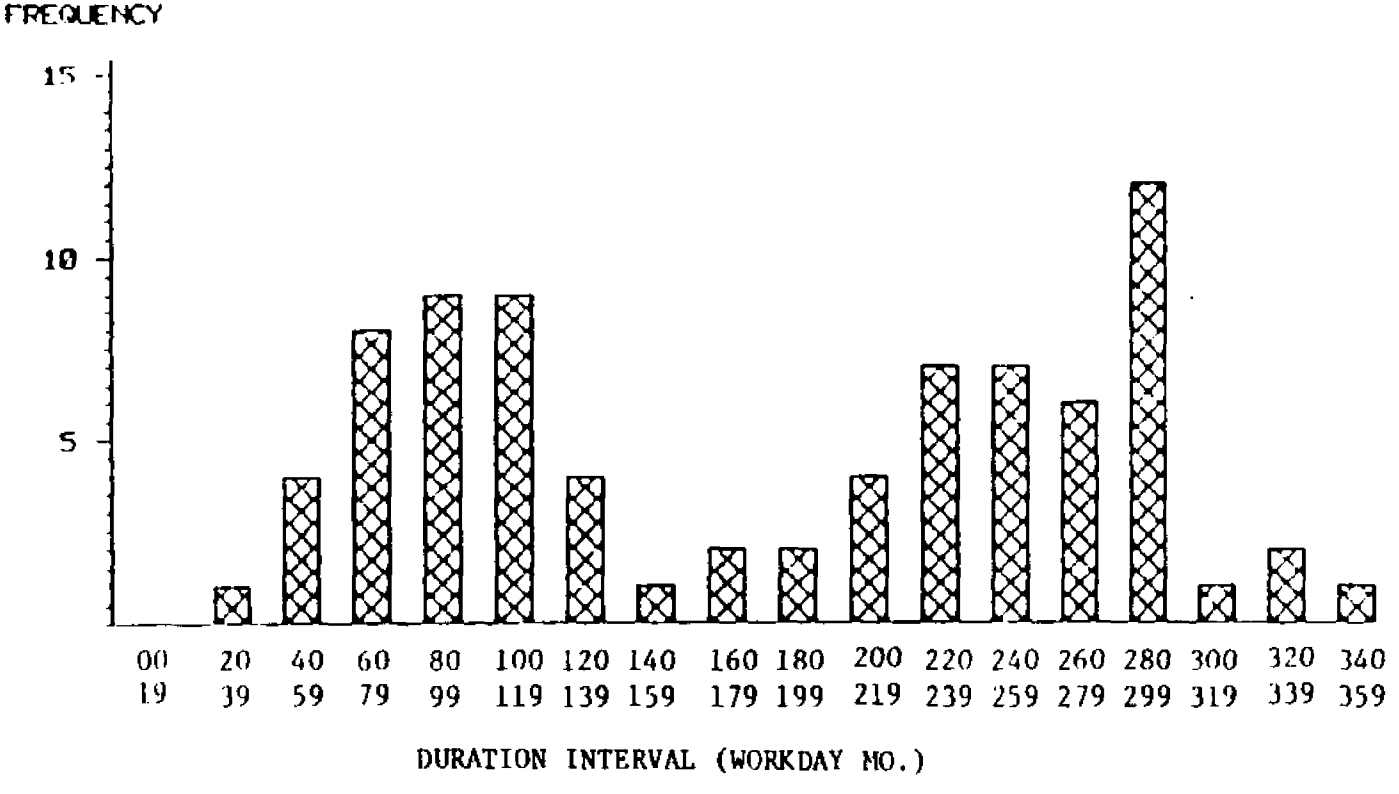


Figure 10. Continued

B. Females

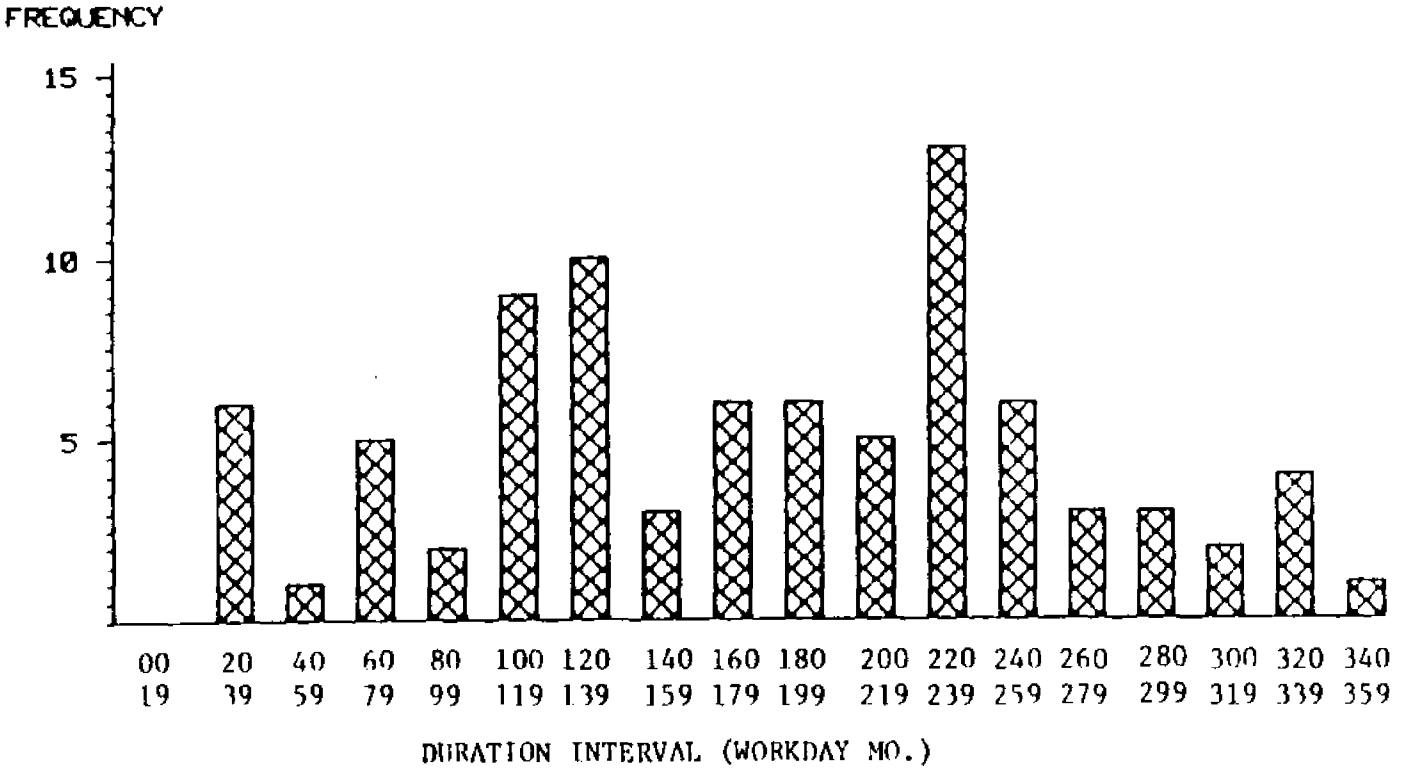


Figure 11. Properties of the job diversity index

Example #	N	k	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}	JDI
1	100	10	10	10	10	10	10	10	10	10	10	10	230
2	100	10	91	1	1	1	1	1	1	1	1	1	0.5
3	100	1	100	0	0	0	0	0	0	0	0	0	0
4	100	100	(n_1 to n_{10} = 1)										460
5	10	10	1	1	1	1	1	1	1	1	1	1	230
6	10	1	10	0	0	0	0	0	0	0	0	0	0

Note the following:

1) #1 vs. #2 - Both have the same number of job categorical changes but different JDI values. #2 has a much less complex job categorical pattern with most of the time spent within one job category.

2) #1 vs. #5 - Both have the same JDI values because the categorical patterns are identical. #1 and #5 would likely make similar recall errors relative to their recall periods.

3) #3 vs. #6 - Both have zero JDI values. A single job category of any duration should be recalled quite accurately.

4) #4 vs. #5 - Both JDI values are maximum for their N-values. #4 has a higher JDI since their is room for more job categorical changes.

5) #1 vs. #4 - #4 has a higher JDI because there are more job categorical changes.

Figure 12. JDI distributions for the original main group

A. Males

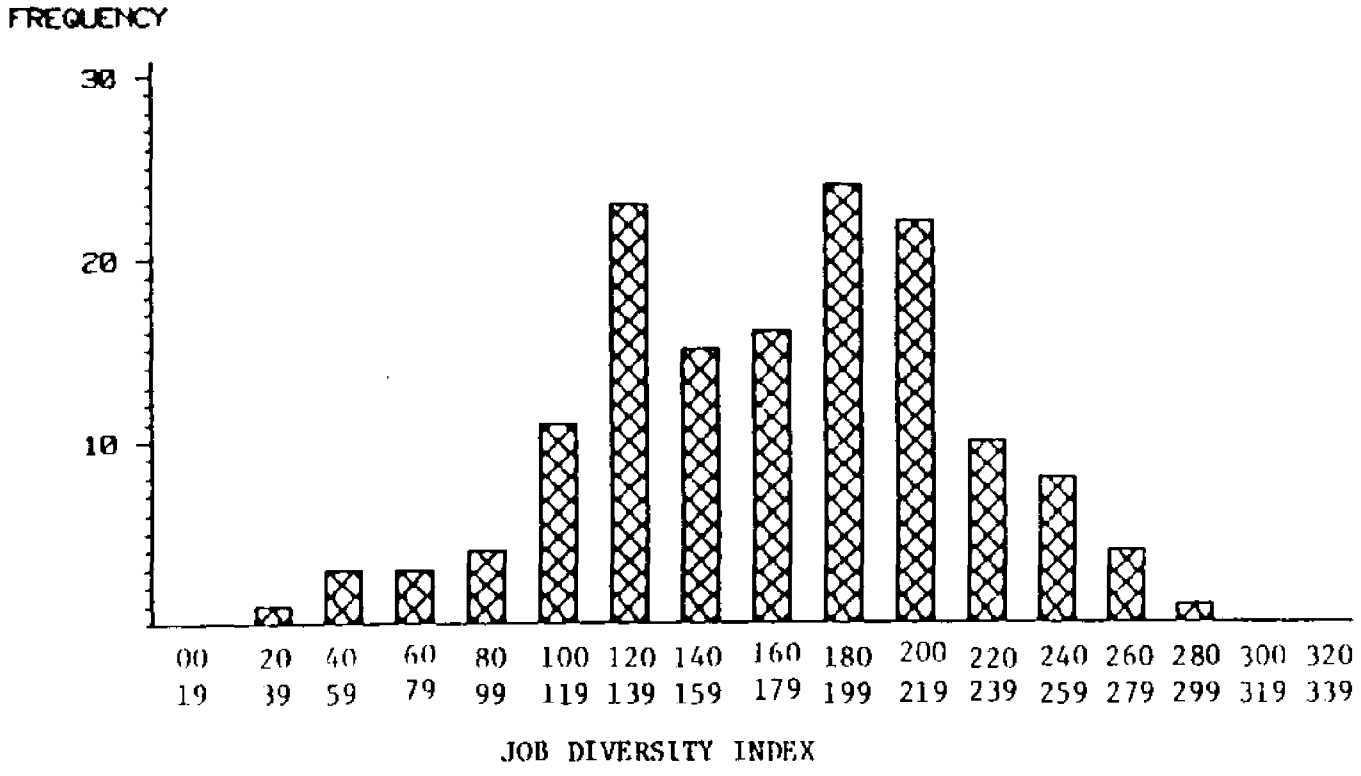


Figure 12. Continued

B. Females

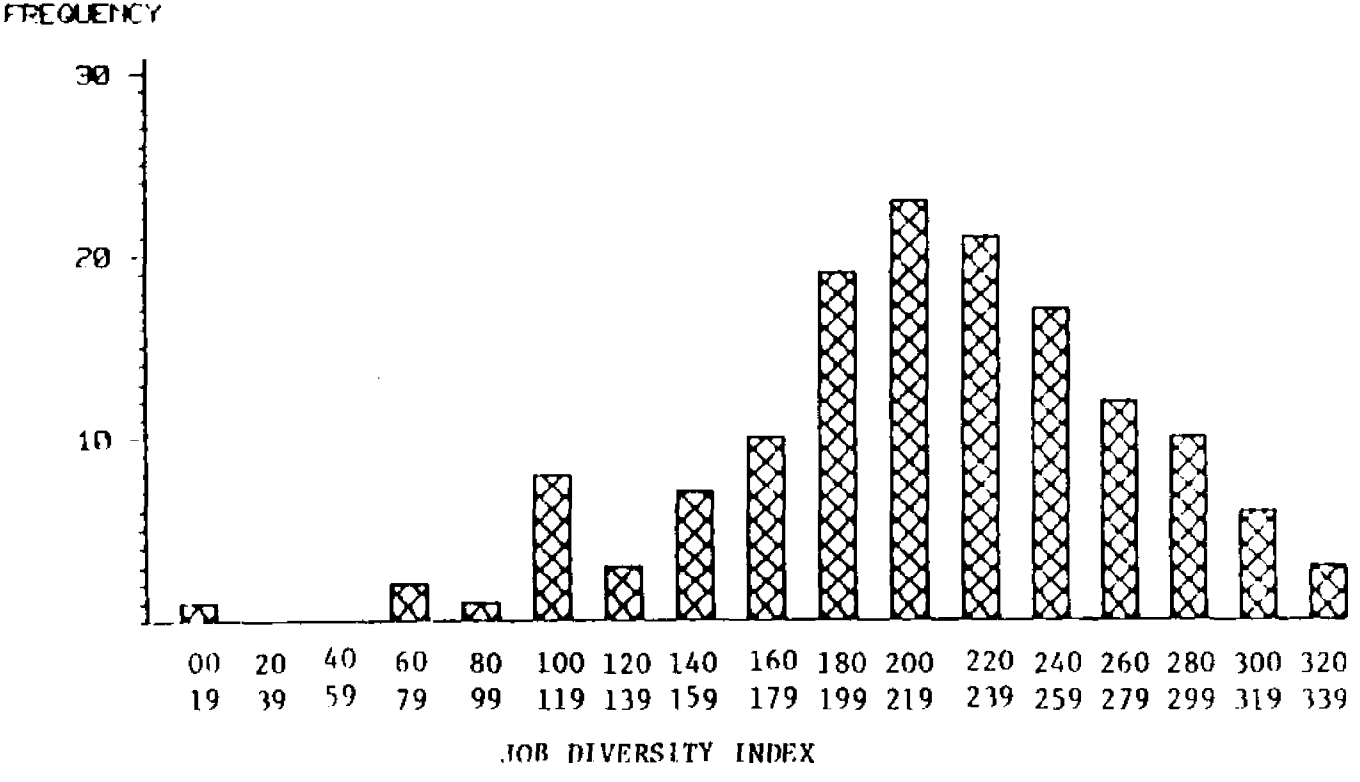


Figure 13. JDI distributions for the reexamined subgroup

A. Males

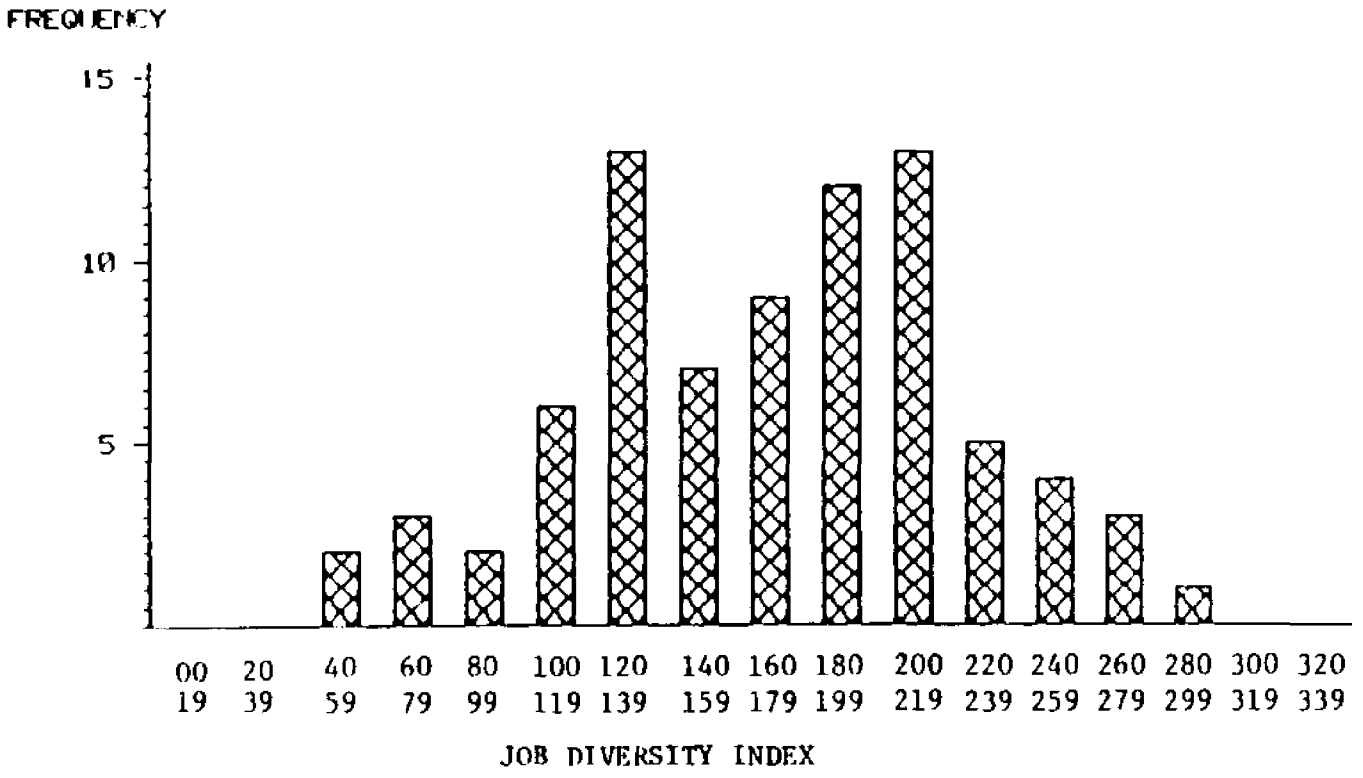
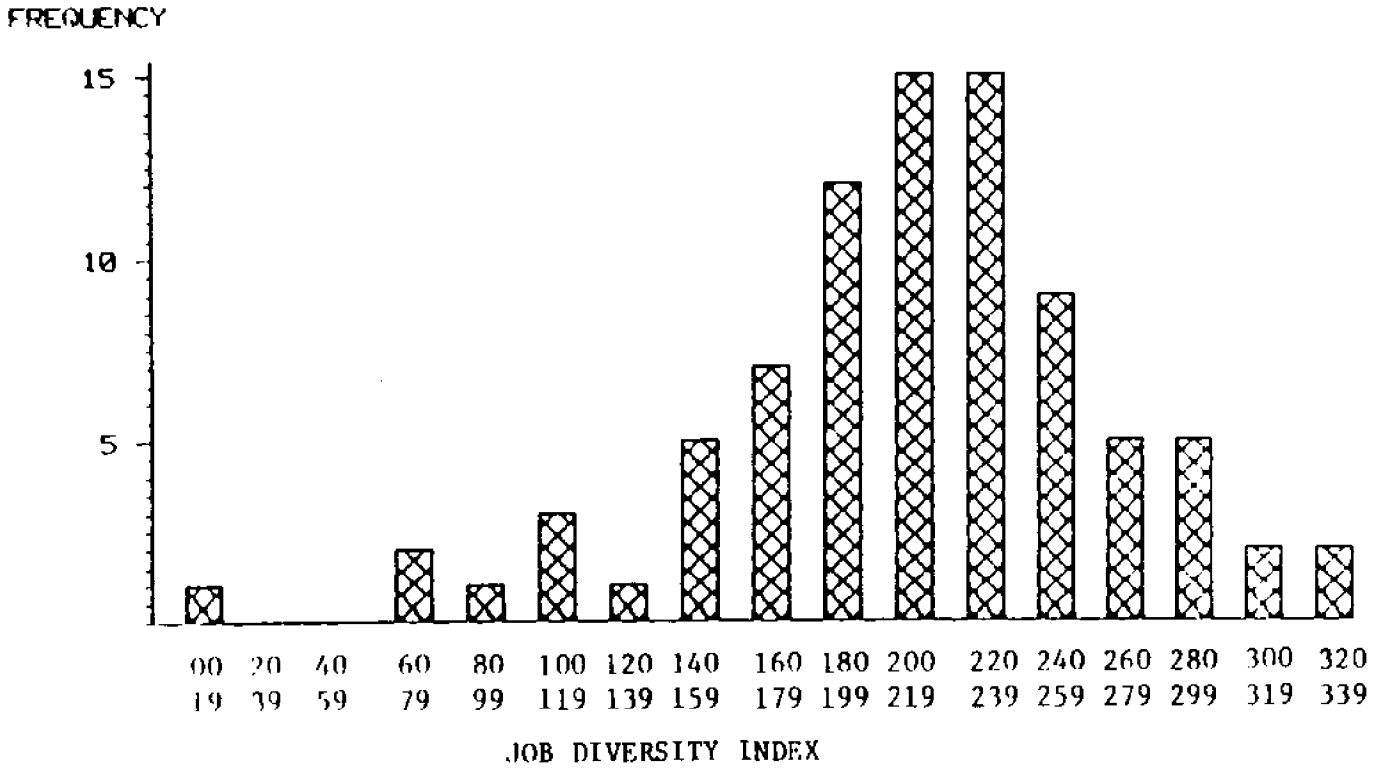


Figure 13. Continued

B. Females



CHAPTER IV: ANALYSIS OF VALIDITY

The following chapter presents the results of the validity procedure. This procedure consisted of two different phases. One of these, the descriptive phase, was concerned with the inaccuracy of self-reporting in this specific study group and the possible misclassifications in individual PCB exposure categories that could have arisen from it (i.e. a specific test of the accuracy of the surveys of Fischbein et al.). The degree of specific accurate reporting was expressed by crude validity scores - that is, scores unadjusted for the effects of independent (possibly confounding) variables, since what the scores would have been in the absence of confounders, or after correction for them, was not what would have determined actual misclassifications. In this phase the crude scores were first presented descriptively, in terms of central tendency and distributional patterns. Then the PCB exposure categories to which the workers were assigned by their self-reported occupational histories were compared to the actual categories (categorical validity) to determine misclassification. Finally, misclassification rates were compared to the crude score results. In the analytical phase, the original crude scores underwent the process of adjustment. The idea was to gain insight into the generalized accuracy of self-reported retrospective occupational histories. In practical terms, the crude validity scores were corrected for the effects of the measureable individual demographic and occupational variables described in the first analysis chapter, producing accuracy profiles unfettered by factors and/or factor levels peculiar to the members of this particular study group (and the population from which they were derived).

INITIAL MULTIPLE REGRESSION ANALYSIS - ORIGINAL MAIN GROUP

The purpose of this first regression analysis was to assess the effects of three specific measureable independent variables, possibly confounders, on accuracy which would have been encountered in any retrospective occupational setting. These variables were the respondent characteristics of: duration of employment (tied up with age), job diversity index, and sex. These were described earlier in the first analysis chapter and were defined according to the company employment records (the validity base). Validity was set up as functions of these variables in the following model:

$$VAL^* = \beta_0 + \beta_1(JDI) + \beta_2(DUR) + \beta_3(SEX) + \beta_4(JDI)(SEX) + \beta_5(DUR)(SEX) + \varepsilon$$

where: DEPENDENT VARIABLE

VAL = validity

* = arcsin square-root transformation

INDEPENDENT VARIABLES

DUR = duration of employment

JDI = job diversity index

SEX = 1 if male, 0 otherwise

$\beta_{(0 \text{ to } 5)}$ = regression coefficients

This particular regression model permitted relation of the response (VAL) to both continuous (DUR, JDI) and class (SEX) independent variables in the same model. Three types of relationships could be

investigated: 1) the response versus the continuous variables (adjusted for the class variables), 2) the effect of the class variable on the response adjusted for the continuous variables, and 3) possible interactions between the continuous and class variables (the final two terms). The model format indicated a linear relationship between the dependent and independent variables. Scatter plots, however, showed that the relationships between both JDI and DUR and VAL were non-linear (Figs. 14 and 15). This was due to the fact that VAL was recorded as a proportion. An appropriate transformation for such a situation is the "arcsin square-root" adjustment. When applied (through SAS procedure), this transformation linearized the data allowing for the use of the selected models and thus linear regression techniques. Additional transformations were unnecessary.

The above model should be thought of as a full linear model. That is, it related the response, VAL*, to a maximum number of possible independent, individual and interaction, effects (β-terms). This full model was fitted to the data initially. To determine if the inclusion of all the β-terms was justified, reduced models, i.e. without various β-terms (those which might only have marginal importance), were fitted. The magnitude of a calculated F-statistic (mean square drop from the reduced to the full model divided by the mean square error of the full model) determined if such terms should have been included. Small F-values suggest elimination. Depending upon the size of the full model, the above process would have to be repeated various numbers of times. When only one β-term has to be considered, the t-value and significance level of the estimated parameter would be sufficient to test for inclusion or elimination. The ultimate aim was to find the most appropriate and

practical model - with all terms significant at at least the .05 level. As a preliminary and guide to the above procedure the SAS technique of STEPWISE-backward elimination was used. This process starts with full models and by eliminating the least significant terms (one by one, in order) produces reduced models containing only those terms significant at the .05 level.

Table 5 gives the results of the initial regression analysis using the aforementioned transformed data. Only those terms whose effects were significant at at least the .05 level were listed in this table or included in the resultant reduced model. The set of two separate equations was derived from the reduced model above them and represented the expected values of VAL* for either males (SEX=1) or females (SEX=0). In the reduced model, the β_0 regression coefficient represented the y-intercept. The β_1 and β_2 terms denoted regression effects (slope) due to independent JDI and DUR effects upon VAL*. The β_3 term stood for an interaction effect caused by the class variable, SEX, which could have either added to or subtracted from the JDI regression effect.

It was evident from the significance levels of the independent variable effects that the mean values of VAL* were not constants, but rather linear functions of the continuous variables, JDI and DUR. Of these two, JDI was the most significant predictor of VAL*. The estimated regression coefficient showed its effect on VAL* to be negative, i.e. it reduced validity as it increased. The DUR was not as highly significant, but was curiously positively correlated with the dependent variable. Most interesting was the significant interaction of the class variable, SEX, with JDI. Fig. 16 shows this effect graphically. At low levels of JDI, the means of male and female validity were both at

or near their maximum values and essentially the same (They had identical y-intercepts). However, as JDI increased in magnitude, the male and female mean values diverged, both decreasing, but the latter at a slower rate (different slopes: |m-males| > |m-females|). Referring back to Table 5, when SEX=1 (male), the JDI regression effect was relatively more negative. In effect, the females, at equivalent JDI levels, had equal or better validity scores than did the males.

In interpreting Fig. 16 one should observe that the y-axis represented the percent validity scale (to show the regression and interaction effects realistically) and that the regression lines were not to be extrapolated beyond their termination points (JDI=280). Also, for the purpose of legibility, only one set of possible regression lines was shown in each figure, specifically the one for DUR=170 months (arbitrary). These sets of regression lines could have shifted either up or down relative to the ones illustrated depending upon the DUR value chosen.

The R-square value for the original main group was 0.395. This represented the percentage of variability that was explained by the chosen model and its constituent measureable independent variables.

TREND ILLUSTRATION - ORIGINAL MAIN GROUP

The purpose of this trend illustration was to demonstrate in tabular form what the regression analysis showed graphically. All predicted mean VAL* values were derived from the male and female regression equations and of course had error terms associated with them. These error terms were not shown, though, as this presentation was meant to be descriptive in nature only.

Table 6 lists the results of this procedure. There was no male-female difference at the lowest JDI level. Both predicted mean values were near maximum, hovering around the 90% validity level. At the next JDI level, 120, predicted mean scores were lower for each sex, but there was still no appreciable sex difference. The female predicted score of 92.0 was only 2.8 points higher than the 89.2 predicted for males. The decrease in predicted mean scores became evident at the third JDI level, 200. This level of complexity was also sufficient to produce a male-female gap of 6.9 points. The predicted male and female scores were 71.9 and 78.8, respectively. At the highest JDI level, the predicted male and female scores dropped to 50.3 and 61.4, respectively. Note that the decrease with JDI became more severe at higher JDI levels. The same was true of the sex effect. At this level, the predicted male-female gap was 11.1 points.

Once again, at the same JDI level, females had either equal or better validity scores than did males (what Fig. 16 showed graphically). Specifically, when JDI was low, validity was high, with no advantage for either sex. At higher JDI levels, though, females exerted a clear validity advantage over males.

CRUDE VALIDITY - ORIGINAL MAIN GROUP

This analysis involved local effects in the original main group. An opening point to be made was that the self-reported occupational histories of this group were not exactly the equivalent of the actual work histories (company employment records). Rather, as Fig. 17 illustrates, the crude validity scores, which measured the degree of accuracy, ranged anywhere from 20 to 99% for males and from 0 to 100% for females. Note,

however, the negatively skewed nature of the validity score distributions, more so for the females, with a concentration of scores in the upper regions of the scale (right) and relatively few in the tail stretching to the lower boundary (left). Thus, while not duplicating the actual work histories, many of the self-reported versions were at least two-thirds accurate (i.e. Self-reporting was not a totally haphazard procedure.). This last assessment, though, was purely descriptive in nature and was not meant to suggest that 70%, or even 80%, accuracy was enough to, in practical terms, prevent serious misclassification. Note that the frequency histograms were constructed with five-point score intervals.

For the presentation of crude validity scores in terms of central tendency, refer to Table 7. Both male and female mean scores were roughly located in the region of 75 to 80% degree of accuracy. Unlike with the JDI-adjusted scores, the male scores here were higher than those for females. The skewness values of -0.90 and -1.31 for males and females, respectively, indicated that an adjusted mean (computed from arcsin square-root transformed values and reconverted to percent equivalent) was the preferred measure for testing the possible significant difference between the male and female scores. The adjusted means were tested, but the difference was not significant (Student's t-test). It must be noted, however, that "no significant difference" still was a departure from the female advantage seen in the initial regression analysis.

SECOND MULTIPLE REGRESSION ANALYSIS - REEXAMINED SUBGROUP

One procedure permitted by the 1979 reexamination was an analysis of the effect of examinational delay on accuracy. The delay tested in

this study was 45 months (the gap between the two evaluation periods - March 1976 to December 1979). The idea was to see if the later a self-reported history was taken (1976 vs. 1979) concerning a specific period, the lower would be its degree of accuracy. Only the reexamined subgroup scores were used in this analysis (as opposed to all the 1976 scores) to insure a comparison within a homogenous population (i.e. Any observed difference in validity would not have been due to differences between the sub- and main groups.). The full linear model in this case was as follows:

$$VAL^* = \beta_0 + \beta_1(JDI) + \beta_2(DUR) + \beta_3(EXP) + \beta_4(JDI)(SEX) + \beta_5(EXP)(JDI) + \beta_6(EXP)(SEX) + \beta_7(JDI)(EXP)(SEX) + \varepsilon$$

where: DEPENDENT VARIABLE

VAL = validity

* = arcsin square-root transformation

INDEPENDENT VARIABLES

DUR = duration of employment

JDI = job diversity index

SEX = 1 if male, 0 otherwise

EXP = 1 if 1976 examination, 0 otherwise

$\beta_{(0 \text{ to } 7)}$ = regression coefficients

Examinational delay was entered in the above full linear model as the class variable, EXP. β-terms found to be insignificant in the initial regression analysis were also eliminated here, i.e. individual sex

effect and DUR interaction. Newly entered were several possible interactions involving EXP. The analytical procedure (arcsin square-root transformation, table and figure format) was the same as in the initial regression analysis.

Table 8 gives the results of the examinational delay regression analysis. Once again, only those terms whose effects were significant at at least the .05 level were listed in this table or included in the resultant reduced model. The set of four separate regression equations represented the following situations: 1) males VAL_{76}* (SEX=1, EXP=1), 2) females VAL_{76}* (SEX=0, EXP=1), 3) males VAL_{79}* (SEX=1, EXP=0), and 4) females VAL_{79}* (SEX=0, EXP=0). In the reduced model, the $\beta_{(0 \text{ to } 2)}$ terms again represented the y-intercept and JDI and DUR regression effects, respectively. The β_3 term represented an intercept effect caused by EXP. The β_4 term represented the interactive JDI*SEX effect seen earlier.

Note first that the JDI, DUR, and interactive JDI*SEX effects seen in the initial analysis were intact for each examinational period (Fig. 18). Considering the effect of examinational delay, EXP, it differed in some respects from that of SEX in the earlier analysis. At low JDI values (<60), the means for 1976 and 1979 VAL* were both at or near their maximum values and virtually equal - so far analagous to earlier SEX effect. With increasing JDI, the 1976 and 1979 lines diverged, as with SEX earlier, reaching a 5 to 6 point difference at JDI=160 (VAL_{76}* > VAL_{79}*). However, unlike with SEX, above the 160 level, this difference remained more or less constant. Returning to Table 8, it can be seen that the EXP effect was more or less an additive constant (intercept effect and not an interactive effect - slopes were not signifi-

cantly different). Thus, except at low JDI levels, the later examination produced a constantly lower validity score. There was no apparent difference between the sexes in the EXP effect. The R-square value for the second model was 0.386, nearly identical to that seen earlier.

TREND ILLUSTRATION - REEXAMINED SUBGROUP

The same type of trend illustration procedure (purely descriptive) was done for the reexamined subgroup. This time there were two factors to be watched at each JDI level, examinational delay and sex. Consistency of effect between the main and subgroups was also important to consider. The results are summarized in Table 9. Note the lack of both sex and examinational delay effects at the lowest JDI level. All of the predicted mean scores, regardless of source were in the high 90% range. Moving to the next level, the sex and examinational delay effects began to emerge. All scores started to decrease, but more so for the males and later examination. The male drops were to 88.6 (1976) and 84.6 (1979) while the corresponding drops for females were only to 92.2 and 88.6, respectively. At the JDI=200 level, the predicted mean scores continued their decline, with the examinational delay effect having reached its constant predicted 5 to 6 point difference (both for males and females). The 1976 predicted scores were 68.9 and 77.7 for males and females, respectively. For 1979, the respective scores were 63.2 and 72.5. At the highest JDI level the scores continued to fall, but with the EXP difference remaining constant. The final predicted scores for the males were 44.6 (1976) and 38.7 (1979). For females the 1976 and 1979 scores were 58.5 and 52.6, respectively.

CRUDE VALIDITY FOR 1976 EXAMINATION - REEXAMINED VS. NON-REEXAMINED SUBGROUP

The carrying out of a reexamination in 1979 split the original group into two subgroups - one that consisted of reexamined individuals and another which was lost to the study after initial examination. One procedure this made possible was a local investigation as to whether or not those individuals who submitted to reexamination presented a different accuracy profile than those who did not. The reexamined subgroup included 165 individuals (57.3%) of the original group. The male-female split in this subgroup was 80-85. The lost subgroup was composed of 123 individuals, 65 of whom were males and the remaining 58, females.

The crude validity scores of the two subgroups were compared in terms of central tendency in Table 10. The negative skewness values again indicated a preference for using adjusted mean values. For males, the adjusted mean difference from the lost to the reexamined subgroup was +1.8 points while the corresponding difference for females was +4.3 points. Crude validity thus appeared to be somewhat higher in the subgroup that opted for reexamination. The difference, however, was not significant (Student's t-test on arcsin square-root adjusted means). Frequency distributions of crude 1976 validity scores for each subgroup are shown in Figs. 19 and 20. Each of the four distributions had the same basic shape as seen earlier with the main group (i.e. negative skew, more so for females; concentration of scores near the upper boundary; and tail regions starting at the two-thirds accuracy level extending back to the lower boundary).

CRUDE VALIDITY FOR REEXAMINED SUBGROUP - 1976 VS. 1979

This analysis involved the local effects of the 45 month examinational delay. Table 11 lists both the 1976 and 1979 crude mean validity scores for the reexamined subgroup. For both males and females the 1979 scores were clearly lower than those for 1976. Adjusted mean crude male validity was 79.6 in 1976, falling to 75.8 in 1979. For females, the adjusted mean crude score was 78.0, dropping to 71.0 in 1979 (The skewness values again dictated the use of adjusted mean values.). In effect, crude validity scores in this subgroup fell, on average, more than 5 points in 45 months. These differences were significant for both males and females (paired t-tests: $p < .0089$ and .0004 for male and female adjusted means, respectively). Examinational delay thus appeared to be an important factor affecting local accuracy. The comparative frequency distributions between 1976 and 1979 are shown in Figs. 21 and 22. Focusing on the upper quartile range of scores (75-100%), note in the male distributions the relatively lower number of scores in this region for 1979. For the females, one could see that this upper quartile region was considerably flatter in 1979 relative to 1976 along with a more consistent and extended tail region. The 1979 female validity distribution also continued to be more negatively skewed than that for the males. The respective values were -1.22 and -0.72, respectively, quite close to the corresponding 1976 values.

THIRD MULTIPLE REGRESSION ANALYSIS - EFFECT OF INTRINSIC TIME LAPSE (THREE SUBPERIOD SUBGROUP)

Up until this point, validity has been treated as if it were constant over the entire recall period. However, the effect of examina-

tional delay on local accuracy (5 point decrease over 45 month period) indicated that degree of validity might be time dependent. For instance, it may depend upon which part of the work history - early, middle, late - one was investigating. Therefore, the initial model was augmented to yield adjusted (by JDI and DUR as earlier) mean 1976 VAL* subscores for the following subrecall decade subperiods: 1950-59, 1960-69, and 1970-76 (i.e. to investigate the effect of intrinsic time lapse on validity). The 1950s decade was selected as the initial subperiod on account of the fact that few individual histories extended back before this era. The full linear model for this analysis was as follows:

$$\begin{aligned} VAL^* = {} & \beta_0 + \beta_1(JDI) + \beta_2(DUR) + \beta_3(SP2) + \beta_4(SP3) + \\ & \beta_5(JDI)(SEX) + \beta_6(JDI)(SP2) + \beta_7(JDI)(SP3) + \\ & \beta_8(SP2)(SEX) + \beta_9(SP3)(SEX) + \beta_{10}(JDI)(SP2)(SEX) + \\ & \beta_{11}(JDI)(SP3)(SEX) + \varepsilon \end{aligned}$$

where: DEPENDENT VARIABLE

VAL = validity

* = arcsin square-root transformation

INDEPENDENT VARIABLES

DUR = duration of employment

JDI = job diversity index

SEX = 1 if male, 0 otherwise

SP2 and SP3 = 0 if 1970s decade subperiod

SP2 = 1 and SP3 = 0 if 1960s decade subperiod

SP2 = 0 and SP3 = 1 if 1950s decade subperiod

$\beta_{(0 \text{ to } 11)}$ = regression coefficients

Intrinsic time lapse was denoted by the class variable, SP. As before, variables found to be insignificant in the initial regression procedure were eliminated here. Possible interaction terms, specific for this analysis, were entered in the above full model (β_6 to β_{10} terms). The analytical procedure was the same as in the initial regression analysis.

Table 12 lists the results for the intrinsic time lapse analysis. One should note the relatively smaller sample sizes in comparison to those seen earlier, since this analysis was restricted to those who worked in all three subperiods. Once again, the reason was to insure a homogenous group so that any observed intrinsic time lapse effects would not have been confounded by differences between older and more recently hired workers. Of course, with this being a restricted or selected subgroup, these individuals might have possessed a differing validity profile from that of the main group. However, it was the possible effect of time lapse that was more of interest here than specific mean scores. Note also the difference between intrinsic time lapse and examinational delay. The latter had to do with a group effect created by the investigator (i.e. How long after the work history occurred was an examination conducted?) while this procedure looked at an effect that was specific to the individual. As before, once again, only those terms significant at at least the .05 level were listed in this table or included in the reduced model. The set of six separate regression equations represented the following situations: 1) males VAL* 1970s SEX=1, SP2=0, SP3=0), 2) females VAL* 1970s (SEX=0, SP2=0, SP3=0), 3)

males VAL* 1960s (SEX=1, SP2=1, SP3=0), 4) females VAL* 1960s (SEX=0, SP2=1, SP3=0), 5) males VAL* 1950s (SEX=1, SP2=0, SP3=1), and 6) females VAL* 1950s (SEX=0, SP2=0, SP3=1). In the reduced model, the β_0 and β_1 terms represented the y-intercept and JDI regression effect, respectively. Note that DUR was eliminated due to the homogeneity of this subgroup in terms of this variable. The β_2 and β_3 terms represented interactive effects caused by the class variable SP. The β_4 term denoted a complex interaction between JDI, SP, and SEX.

An initial observation was that the JDI effect was operative as in both previous regression analyses (Fig. 23). However, its magnitude varied significantly with the variable SP. The SP effect was interactive - At low JDI values (<40) there was no effect. All scores were at the maximum level. Sufficient increases in JDI yielded generally decreased validity, but more so with earlier subperiod (different slopes: |m1950s| > |m1960s| > |m1970s|). The lines continued to diverge with further JDI increases, except that the male 1950s line began to level off as it approached its minimum value. Note that the sex effect was also dependent upon SP. There was no significant sex effect for the most recent subperiod while it was evident in the two earlier ones. The sex effect was not significantly different between SP2 and SP3, hence the common β-term.

The implication in these results was that the time lapse effect was both significant and a JDI dependent phenomenon. The overall validity scores were composites of three significantly different subscores. Only when JDI was sufficiently high did any time lapse trends appear. Conversely, JDI was not as important in effect until there was sufficient time lapse (Note the small decrease with JDI for the 1970s sub-

period). In essence there was a mutually dependent relationship. As to sex, its effect was not only dependent upon JDI but upon intrinsic time lapse. Females were at least equal to males (low JDI, recent subperiod) and increasingly superior with increased JDI and earlier subperiod, although the gap narrowed as the scores approached their minimum levels. The R-square value for this analysis was 0.431, higher than seen earlier (Note that this was a more homogenous group.).

TREND ILLUSTRATION FOR INTRINSIC TIME LAPSE ANALYSIS - THREE SUBPERIOD SUBGROUP

This trend illustration added the factor of intrinsic time lapse to the original JDI trend. The results are contained within Table 13. Most apparent was that there was little or no intrinsic time lapse effect or sex difference at the lowest JDI level. All of the six predicted mean scores in this group were virtually 100%. Also, no matter what the JDI level, the mean values for the most recent period, 1970-76, were near maximum (lowest was 84.6) and equivalent for the sexes. At the JDI=200 level, both the time lapse lapse and sex effects surfaced. The 90+% predicted mean scores of the 1970s at this JDI level fell to 55.4 for males, but only to 69.0 for the females, in the 1950s. At the highest JDI level, the time lapse effect was even greater. Here, the predicted mean drop was to 40.7 for females and 22.6 for males. Note again the levelling off of the males 1950s predicted mean score.

CRUDE VALIDITY - DECADE SUBSCORES (THREE SUBPERIOD SUBGROUP)

The local effects of intrinsic time lapse for the 1976 examination are shown in Table 14. Looking at the results, there was an obvi-

ous decrease in adjusted mean crude validity the earlier one went back in the history. For both males and females there was no problem recalling the most recent period of the work history as mean scores for 1970-76 were at or near the 90% range. This situation changed dramatically moving backward in time. The scores dropped significantly in the 1960s decade and continued falling significantly to the 60% range for the earliest period (ANOVA - randomized block, Duncan's new multiple range test on arcsin square-root adjusted means). Overall, the mean declines with time lapse for each sex were roughly equivalent. For males the change was -31.8 points and for females, -25.2 points.

EVIDENCE OF MISCLASSIFICATION

The purpose of this analysis was to determine what degree of, if any, misclassification was produced in this study group through the use of self-reported occupational histories. Then, if significant levels did exist, the idea was to see how they were related to the observed local degree of non-validity. According to the agenda laid out in the methods section, each individual was assigned two overall total PCB risk figures, one for each of the two PCB subperiods (HPCB: 1947-71, LPCB: 1971-76). Each risk figure had at least two versions, one according to company employment records (actual) and another according to the 1976 occupational history (self-reported). A third version, also self-reported (1979 history), appeared for the members of the reexamined subgroup. Each overall total risk figure ultimately assigned the worker to one of four exposure categories. These categories were based upon quartiles from the distribution of actual total overall PCB risk figures. The limits of these categories are shown in Table 15 and were

quite broad. Thus, many shifts from one category to the next could have represented substantial changes in estimated exposure.

The misclassification results appear in Table 16. Only the re-examined subgroup was used in this analysis to allow for unbiased 1976-1979 comparisons. All shifts in exposure category were defined as relative to the actual category (company record version). Validity in this table denoted the percentage of individuals whose self-reported exposure categories were consistent with the actual one.

For the 1976 examination (HPCB period), there were 27 total shifts away from the actual category. Some of these shifts spanned more than one category. Note the predominance of upward shifts, 20, versus 7 downward. For the 1979 examination (HPCB period) there were even more shifts, 42. Again, note the greater tendency to shift upward relative to the company record version (31 upward versus 11 downward). This tendency to shift upward (HPCB period) was significant for both examinations (sign test - 1976: $p < .0124$, 1979: $p < .0020$). In essence, there was a tendency for the study subjects to overestimate time spent in high exposure job categories.

The exposure categorical validity for the 1976 examination (HPCB period) was 82.8%. For that of 1979 (HPCB period) it fell to 73.2%. The respective misclassification rates were thus 17.8% and 26.8%. Note the effect of examinational delay on misclassification - a higher rate for delayed examination. The tendency for greater misclassification for the 1979 examination (HPCB period) was significant (McNemar's test, $p < .0050$).

The 17.8% misclassification rate (HPCB period, 1976) corresponded to an adjusted mean crude non-validity of 20-22%, while the 26.8% figure

for 1979 (HPCB period) corresponded to a mean crude non-validity of 24-29% (see Table 11). In terms of consistency, the higher misclassification rate followed the higher mean crude non-validity percentage.

For the LPCB period, the results were similar, although none of the tendencies or differences reached significance. Note, though, that in terms of recall, this was a more recent period.

FOURTH MULTIPLE REGRESSION ANALYSIS - INTERVIEWER EFFECTS

This regression analysis involved the addition of the independent class variable, interviewer, INT, to the initial model type. The idea was to see if interviewer differences in technique, skill, impartiality, perseverence, consistency, case load, etc. could lead to significant differences in mean interviewer-specific accuracy. This potential effect in generalized terms would of course be dependent upon the role given to interviewers by the specific study design (i.e. There could be situations with no interviewers, a single interviewer, two, or more.). In this study, there were nine different interviewers for the initial 1976 examination and seven different for the 1979 reexamination. The dependent variable, VAL* was set up as separate functions for 1976 and 1979 interviewers as follows (full linear model):

$$1)\ VAL^{*} = \beta_0 + \beta_1(JDI) + \beta_2(DUR) + \beta_3(SEX) + \beta_4(INT2) + \beta_5(INT3) + \beta_6(INT4) + \beta_7(INT5) + \beta_8(INT6) + \beta_9(INT7) + \beta_{10}(INT8) + \beta_{11}(INT9) + \beta_{12}(JDI)(SEX) + \beta_{13}(INT2)(SEX) + \beta_{14}(INT3)(SEX) + \beta_{15}(INT4)(SEX) + \beta_{16}(INT5)(SEX) + \beta_{17}(INT6)(SEX) + \beta_{18}(INT7)(SEX) + \beta_{19}(INT8)(SEX) + \beta_{20}(INT9)(SEX) + \beta_{21}(INT2)(JDI) +$$

$$\beta_{22}(INT3)(JDI) + \beta_{23}(INT4)(JDI) + \beta_{24}(INT5)(JDI) +$$
$$\beta_{25}(INT6)(JDI) + \beta_{26}(INT7)(JDI) + \beta_{27}(INT8)(JDI) +$$
$$\beta_{28}(INT9)(JDI) + \beta_{29}(INT2)(JDI)(SEX) +$$
$$\beta_{30}(INT3)(JDI)(SEX) + \beta_{31}(INT4)(JDI)(SEX) +$$
$$\beta_{32}(INT5)(JDI)(SEX) + \beta_{33}(INT6)(JDI)(SEX) +$$
$$\beta_{34}(INT7)(JDI)(SEX) + \beta_{35}(INT8)(JDI)(SEX) +$$
$$\beta_{36}(INT9)(JDI)(SEX) + \varepsilon$$

2) $VAL^* = \beta_0 + \beta_1(JDI) + \beta_2(DUR) + \beta_3(SEX) + \beta_4(INT2') +$
$\beta_5(INT3') + \beta_6(INT4') + \beta_7(INT5') + \beta_8(INT6') +$
$\beta_9(INT7') + \beta_{10}(JDI)(SEX) + \beta_{11}(INT2')(SEX) +$
$\beta_{12}(INT3')(SEX) + \beta_{13}(INT4')(SEX) + \beta_{14}(INT5')(SEX) +$
$\beta_{15}(INT6')(SEX) + \beta_{16}(INT7')(SEX) + \beta_{17}(INT2')(JDI) +$
$\beta_{18}(INT3')(JDI) + \beta_{19}(INT4')(JDI) + \beta_{20}(INT5')(JDI) +$
$\beta_{21}(INT6')(JDI) + \beta_{22}(INT7')(JDI) + \beta_{23}(INT2')(JDI)(SEX) +$
$\beta_{24}(INT3')(JDI)(SEX) + \beta_{25}(INT4')(JDI)(SEX) +$
$\beta_{26}(INT5')(JDI)(SEX) + \beta_{27}(INT6')(JDI)(SEX) +$
$\beta_{28}(INT7')(JDI)(SEX) + \varepsilon$

where: DEPENDENT VARIABLES

VAL = validity

* = arcsin square-root transformation

INDEPENDENT VARIABLES

DUR = duration of employment

JDI = job diversity index

INT2 to 9 = 0 if interviewer 1

INT2 = 1 if interviewer 2, 0 otherwise

INT3 = 1 if interviewer 3, 0 otherwise, etc.

INT2' to 7' = 0 if interviewer 1'

INT2' = 1 if interviewer 2', 0 otherwise

INT3' = 1 if interviewer 3', 0 otherwise, etc.

$\beta_{(0 \text{ to } 36)}$ = regression coefficients

These full linear models were handled as in the earlier analysis (including the transformation). All possible appropriate new interactions were included.

Tables 17 and 18 give the results of this multiple regression analysis. As earlier, only those terms whose effects were significant at at least the .05 level were listed in these tables or included in the reduced models. This time there were two reduced models following the regression procedure - VAL* for 1976 and 1979 interviewers. Each reduced model generated a variable number of expected value equations for individual interviewers. The lack of an equation for some interviewers was due to insufficient numbers of points. In the reduced models the β_0 regression coefficient represented the y-intercept. The β_1 and β_2 terms denoted regression effects (slope) due to independent JDI and DUR effects upon VAL*. The remaining β-terms represented either main interviewer effects which could either have added to or subtracted from the intercept (positional) value, interactions between interviewers and sex of the respondent (sex specific intercept effect which could have added to or subtracted from the intercept for one sex only), or interaction between interviewers and both JDI and sex (sex specific interaction effect which could have either increased or decreased slope effects for

one sex only).

Starting with the results for 1976 (see also Fig. 24), it was evident that the male mean value regression line from Fig. 16 (initial regression analysis) was really a composite (average) of at least two significantly different interviewer lines. Note that INT3 had a significantly higher intercept but yet a more severe negative interaction effect with increasing JDI than did the other interviewers. Below JDI values of 200, this interviewer produced relatively higher validity scores, yet this advantage disappeared and eventually became a disadvantage. At JDI=280, the INT3 predicted mean VAL* was about 15 points less than that for all other interviewers. INT3 effects were restricted to male respondents only. Any interviewer differences when dealing with female respondents could not be distinguished. Note that the negative interactive sex effect on male respondents by INT3 did not account for the overall interactive sex effect. In summary, there were two 1976 interviewer effects which either enhanced or diminished validity relative to the composite value. INT7 and INT9 had too few points to be factors in this analysis. Moving to the R-square value, 0.411, it was higher in this regression analysis than in the original one. Thus, adding interviewer effects into the models accounted for (explained) more of the variability.

For 1979 (see also Fig. 25) there was quite an interesting result. Both male and female validity were significantly influenced by INT5'. The effects were in both cases of the intercept type (i.e. virtually constant difference at most JDI levels - not interactive). However, the effects on male and female respondents were antagonistic. At JDI values greater than 160, relative to the other interviewers, INT5'

produced a constant advantage for males of approximately 8 points while producing about a 25 point disadvantage for females at the same time. All other interviewers could not be distinguished. The overall sex interactive effect seen earlier (female advantage) remained not only intact here, but was made to appear stronger than indicated by the composite due to the advantage given to male respondent validity by INT5'. The R-square for 1979, 0.406, was higher, once again, than that for the initial regression analysis. Assessing the above results, it could be seen that interviewers could exert considerably varying influences upon validity scores.

Table 5. Independent measureable variable effects on 1976 validity

Initial multiple regression analysis (original main group)
N=288, R^2=0.395, F=61.81 ($p < .0001$)

Parameter	Estimate	t for H_0: β=0	Prob. > |t|
Intercept	1.47	39.18	.0001
JDI	-0.0024	-13.36	.0001
DUR	0.0006	5.01	.0001
JDI*SEX	-0.0004	- 4.05	.0001

Reduced model: $VAL^* = \beta_0 + \beta_1(JDI) + \beta_2(DUR) + \beta_3(JDI)(SEX) + \varepsilon$

Regression equations (expected values):

Male (SEX=1): $VAL^* = 1.47 - 0.0028(JDI) + 0.0006(DUR)$

Female (SEX=0): $VAL^* = 1.47 - 0.0024(JDI) + 0.0006(DUR)$

Table 6. Predicted mean VAL* (%) for selected JDI levels (original main group)

JDI Level	Predicted Mean VAL* (%) Males	Females
40	98.7	99.1
120	89.2	92.0
200	71.9	78.8
280	50.3	61.4

Table 7. Crude validity scores (%) for the original main group

Sex	Number	Mean	Std. Dev.	Std. Err.	Adj. Mean[a]	Skewness
M	145	76.6	17.6	1.46	78.8	-0.90
F	143	74.3	18.2	1.52	76.2	-1.30

[a]Male-female difference was NS (Student's t-test on arcsin square-root adjusted means)

Table 8. Independent measureable variable effects on 1976 and 1979 validity

Second multiple regression analysis (reexamined subgroup)
N=165, R^2=0.386, F=50.74 ($p < .0001$)

Parameter	Estimate	t for H_0: β=0	Prob. > |t|
Intercept	1.42	34.31	.0001
JDI	-0.0026	-13.51	.0001
DUR	0.0007	5.84	.0001
EXP	0.06	2.92	.0037
JDI*SEX	-0.0005	- 4.15	.0001

Reduced model: $VAL^* = \beta_0 + \beta_1(JDI) + \beta_2(DUR) + \beta_3(EXP) + \beta_4(JDI)(SEX) + \varepsilon$

Regression equations (expected values):

Male 1976 (SEX=1,EXP=1): $VAL^* = 1.48 - 0.0031(JDI) + 0.0007(DUR)$

Male 1979 (SEX=1,EXP=0): $VAL^* = 1.42 - 0.0031(JDI) + 0.0007(DUR)$

Female 1976 (SEX=0,EXP=1): $VAL^* = 1.48 - 0.0026(JDI) + 0.0007(DUR)$

Female 1979 (SEX=0,EXP=0): $VAL^* = 1.42 - 0.0026(JDI) + 0.0007(DUR)$

Table 9. Predicted mean VAL* (%) for 1976 and 1979 and selected JDI levels (reexamined subgroup)

	JDI = 40		JDI = 120		JDI = 200		JDI = 280	
	Pr. Mean VAL*		Pr. Mean VAL*		Pr. Mean VAL*		Pr. Mean VAL*	
EXP	Male	Female	Male	Female	Male	Female	Male	Female
1976	99.1	99.4	88.6	92.2	68.9	77.7	44.6	58.5
1979	97.6	98.2	84.6	88.6	63.2	72.5	38.7	52.6

Table 10. Crude validity scores (%) for reexamined and non-reexamined subgroups

Male						
Reex.	Number	Mean	Std. Dev.	Std. Err.	Adj. Mean[a]	Skewness
Yes	80	77.2	16.8	1.88	79.6	-0.75
No	65	75.8	17.5	2.17	77.8	-1.08

Female						
Reex.	Number	Mean	Std. Dev.	Std. Err.	Adj. Mean[a]	Skewness
Yes	85	75.8	18.2	1.97	78.0	-1.31
No	58	72.2	18.1	2.38	73.7	-1.36

[a]Reex.-nonreex. differences were NS (Student's t-test on arcsin square-root adjusted means)

Table 11. Crude 1976 and 1979 validity scores (%) for the reexamined subgroup

Male						
EXP	Number	Mean	Std. Dev.	Std. Err.	Adj. Mean[a]	Skewness
1976	80	77.2	16.8	1.88	79.6	-0.75
1979	80	73.5	18.3	2.05	75.8	-0.72

Female						
EXP	Number	Mean	Std. Dev.	Std. Err.	Adj. Mean[a]	Skewness
1976	85	75.8	18.2	1.97	78.0	-1.31
1979	85	69.4	23.2	2.52	71.0	-1.22

[a]1976-1979 male difference was significant (paired t-test on arcsin square-root adjusted means, $p < .0089$)
1976-1979 female difference was significant (paired t-test on arcsin square-root adjusted means, $p < .0004$)

Table 12. Independent measureable variable effects on 1976 subperiod validity

Third multiple regression analysis (three subperiod subgroup)
N=127, R^2=0.431, F=67.31 ($p < .0001$)

Parameter	Estimate	t for H_0: β=0	Prob. > |t|
Intercept	1.70	32.31	.0001
JDI	-0.0019	- 7.42	.0001
JDI*SP2	-0.0006	- 3.39	.0008
JDI*SP3	-0.0017	-11.66	.0001
JDI*SP2or3*SEX	-0.0007	- 3.49	.0005

Reduced model: $VAL^* = \beta_0 + \beta_1(JDI) + \beta_2(JDI)(SP2) + \beta_3(JDI)(SP3) + \beta_4(JDI)(SP2or3)(SEX) + \varepsilon$

Regression equations (expected values):

Male 1970s (SEX=1,SP2=0,SP3=0): $VAL^* = 1.70 - 0.0019(JDI)$

Male 1960s (SEX=1,SP2=1,SP3=0): $VAL^* = 1.70 - 0.0032(JDI)$

Male 1950s (SEX=1,SP2=0,SP3=1): $VAL^* = 1.70 - 0.0043(JDI)$

Female 1970s (SEX=0,SP2=0,SP3=0): $VAL^* = 1.70 - 0.0019(JDI)$

Female 1960s (SEX=0,SP2=1,SP3=0): $VAL^* = 1.70 - 0.0025(JDI)$

Female 1950s (SEX=0,SP2=0,SP3=1): $VAL^* = 1.70 - 0.0036(JDI)$

Table 13. Predicted mean VAL* (%) for decade subperiods and selected JDI levels (three subperiod subgroup)

	JDI = 40		JDI =120		JDI = 200		JDI = 280	
	Pr. Mean VAL*		Pr. Mean VAL*		Pr. Mean VAL*		Pr. Mean VAL*	
SP	Male	Female	Male	Female	Male	Female	Male	Female
1970s	100.0	100.0	99.0	99.0	93.8	93.8	84.6	84.6
1960s	100.0	100.0	93.6	97.1	76.1	86.9	51.9	70.8
1950s	99.8	100.0	85.7	91.1	55.4	69.0	22.6	40.7

Table 14. Crude 1976 validity scores (%) broken down by decade subperiod for the three subperiod subgroup

Male						
SP	Number	Mean	Std. Dev.	Std. Err.	Adj. Mean[a]	Skewness
1970s	57	91.6	16.6	2.20	98.6	-2.33
1960s	57	74.7	22.7	3.01	78.5	-0.64
1950s	57	64.1	24.8	3.28	66.8	-0.54

Female						
SP	Number	Mean	Std. Dev.	Std. Err.	Adj. Mean[a]	Skewness
1970s	70	86.6	13.8	1.65	88.8	-1.55
1960s	70	79.4	24.9	2.98	82.6	-1.73
1950s	70	61.4	27.3	3.26	63.6	-0.43

[a]1970s-1960s and 1960s-1950s male differences were significant (ANOVA - randomized block, Duncan's new multiple range test at $\alpha = .01$ on arcsin square-root adjusted means)
1970s-1960s and 1960s-1950s female differences were significant (ANOVA - randomized block, Duncan's new multiple range test at $\alpha = .05$ on arcsin square-root adjusted means)

Table 15. Limits of PCB exposure categories

Category	HPCB Level	LPCB Level
1	0-74	0-37
2	75-195	38-62
3	196-470	63-198
4	>470	>198

Table 16. Misclassification of PCB exposure categories in the reexamined subgroup

Exposure to HPCBs (up to 1971)	
1976 Examination: N=157	1979 Examination: N=157
Upward Shifts:[a] 20	Upward Shifts:[a] 31
Downward Shifts: 7	Downward Shifts: 11
Total Shifts: 27	Total Shifts: 42
Exp. Cat. Validity: 82.8%	Exp. Cat. Validity: 73.2%
Misclass. Rate:[b] 17.2%	Misclass. Rate:[b] 26.8%

Exposure to LPCBs (1971-1976)	
1976 Examination: N=157	1979 Examination: N=157
Upward Shifts:[c] 16	Upward Shifts:[c] 18
Downward Shifts: 7	Downward Shifts: 15
Total Shifts: 23	Total Shifts: 33
Exp. Cat. Validity: 85.3%	Exp. Cat. Validity: 79.0%
Misclass. Rate:[d] 14.7%	Misclass. Rate:[d] 21.0%

[a] 1976 tendency to shift upward was significant (sign test, $p < .0124$)
1979 tendency to shift upward was significant (sign test, $p < .0020$)

[b] 1976-1979 difference was significant (McNemar's test, $p < .0050$)

[c] 1976 tendency to shift upward was NS (sign test)
1979 tendency to shift upward was NS (sign test)

[d] 1976-1979 difference was NS (McNemar's test)

Table 17. Interviewer effects on 1976 validity

Fourth multiple regression analysis (original main group)
N=288, R^2=0.411, F=39.42 ($p < .0001$)

Parameter	Estimate	t for H_0: β=0	Prob. > |t|
Intercept	1.44	37.22	.0001
JDI	-0.0022	-12.44	.0001
DUR	0.0006	5.17	.0001
JDI*SEX	-0.0005	- 4.03	.0001
INT3*SEX	0.34	2.53	.0119
JDI*INT3*SEX	-0.0017	- 2.09	.0379

Reduced model: $VAL^* = \beta_0 + \beta_1(JDI) + \beta_2(DUR) + \beta_3(JDI)(SEX) + \beta_4(INT3)(SEX) + \beta_5(JDI)(INT3)(SEX) + \varepsilon$

Regression equations (expected values):

Male INT1,2,4,5,6,8: $VAL^* = 1.44 - 0.0027(JDI) + 0.0006(DUR)$

Male INT3: $VAL^* = 1.78 - 0.0044(JDI) + 0.0006(DUR)$

Female INT1,2,3,4,5,6: $VAL^* = 1.44 - 0.0022(JDI) + 0.0006(DUR)$

Table 18. Interviewer effects on 1979 validity

Fourth multiple regression analysis (reexamined subgroup)
N=165, R^2=0.406, F=21.69 ($p < .0001$)

Parameter	Estimate	t for H_0: β=0	Prob. > |t|
Intercept	1.43	22.99	.0001
JDI	-0.0025	- 8.50	.0001
DUR	0.0008	3.97	.0001
INT5'	-0.28	- 3.94	.0001
JDI*SEX	-0.0006	- 3.25	.0014
INT5'*SEX	0.36	3.57	.0005

Reduced model: $VAL^* = \beta_0 + \beta_1(JDI) + \beta_2(DUR) + \beta_3(INT5') + \beta_4(JDI)(SEX) + \beta_5(INT5')(SEX) + \varepsilon$

Regression equations (expected values):

Male INT2',3',4',6': $VAL^* = 1.43 - 0.0031(JDI) + 0.0008(DUR)$

Male INT5': $VAL^* = 1.51 - 0.0031(JDI) + 0.0008(DUR)$

Female INT1',2',3',4',6': $VAL^* = 1.43 - 0.0025(JDI) + 0.0008(DUR)$

Female INT5': $VAL^* = 1.15 - 0.0025(JDI) + 0.0008(DUR)$

Figure 14. Relationship between 1976 validity and job diversity index (original main group)

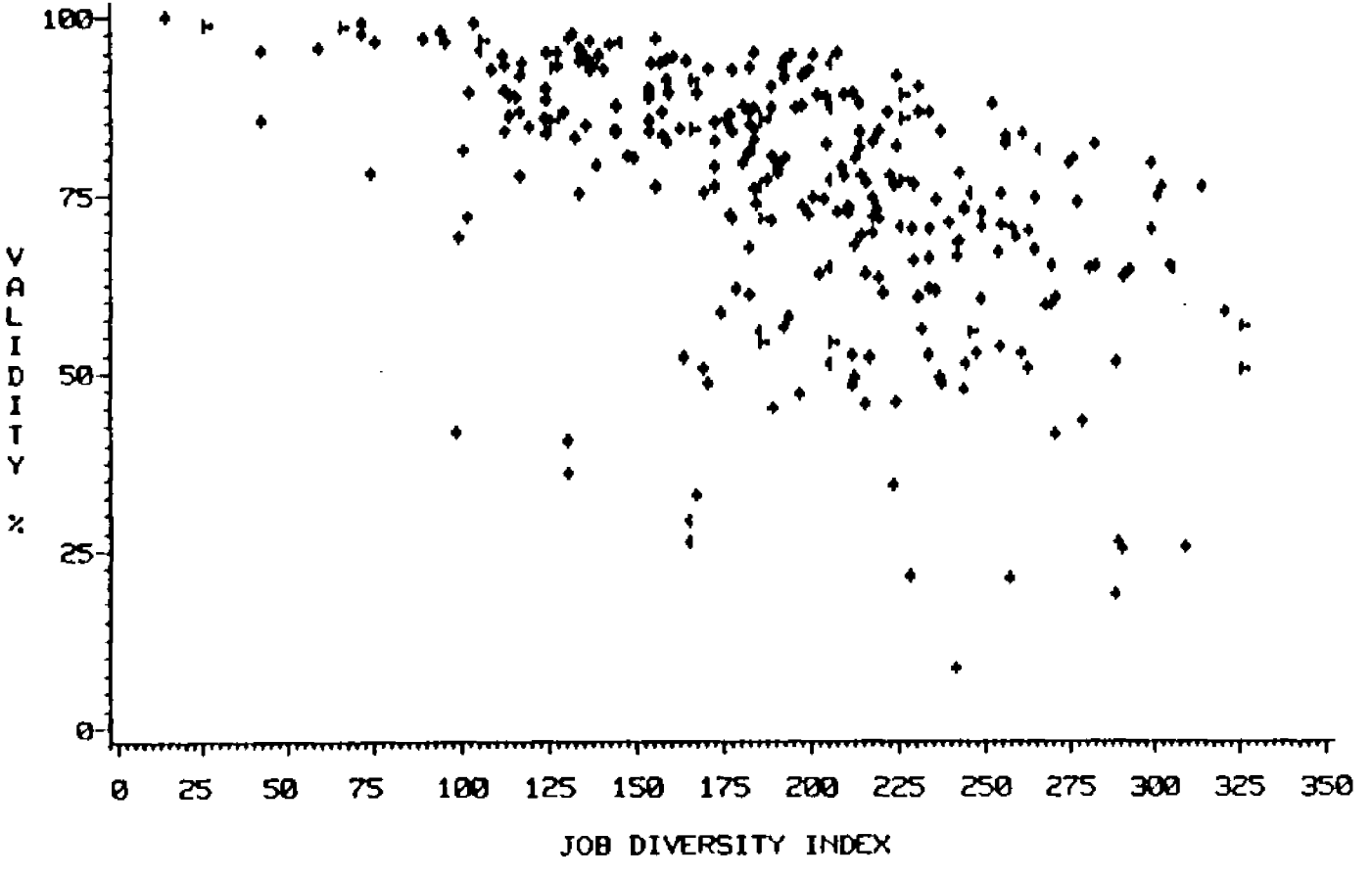


Figure 15. Relationship between 1976 validity and duration of employment (original main group)

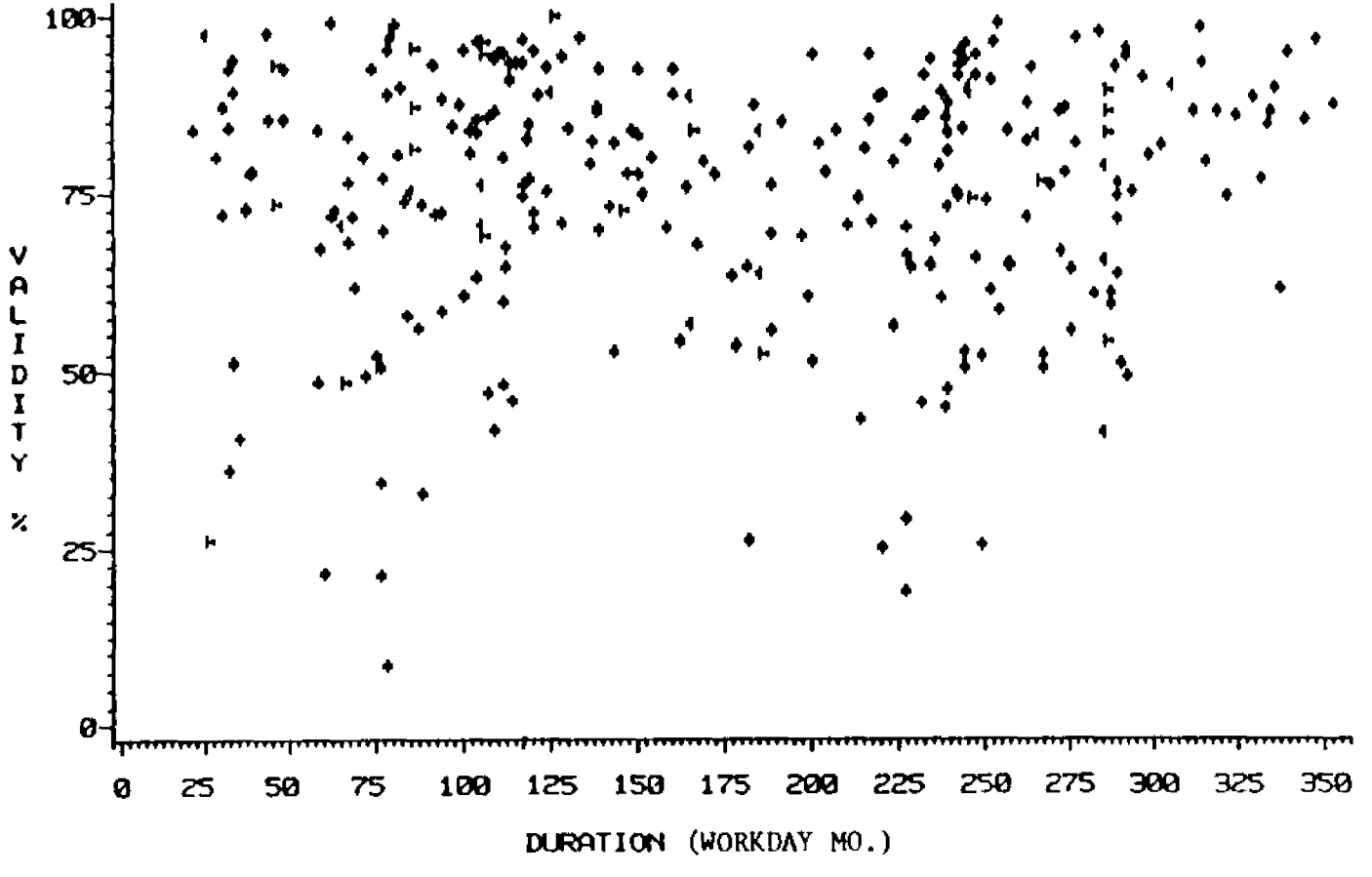


Figure 16. Regression lines describing relationship between 1976 validity and JDI, SEX, and DUR (original main group)

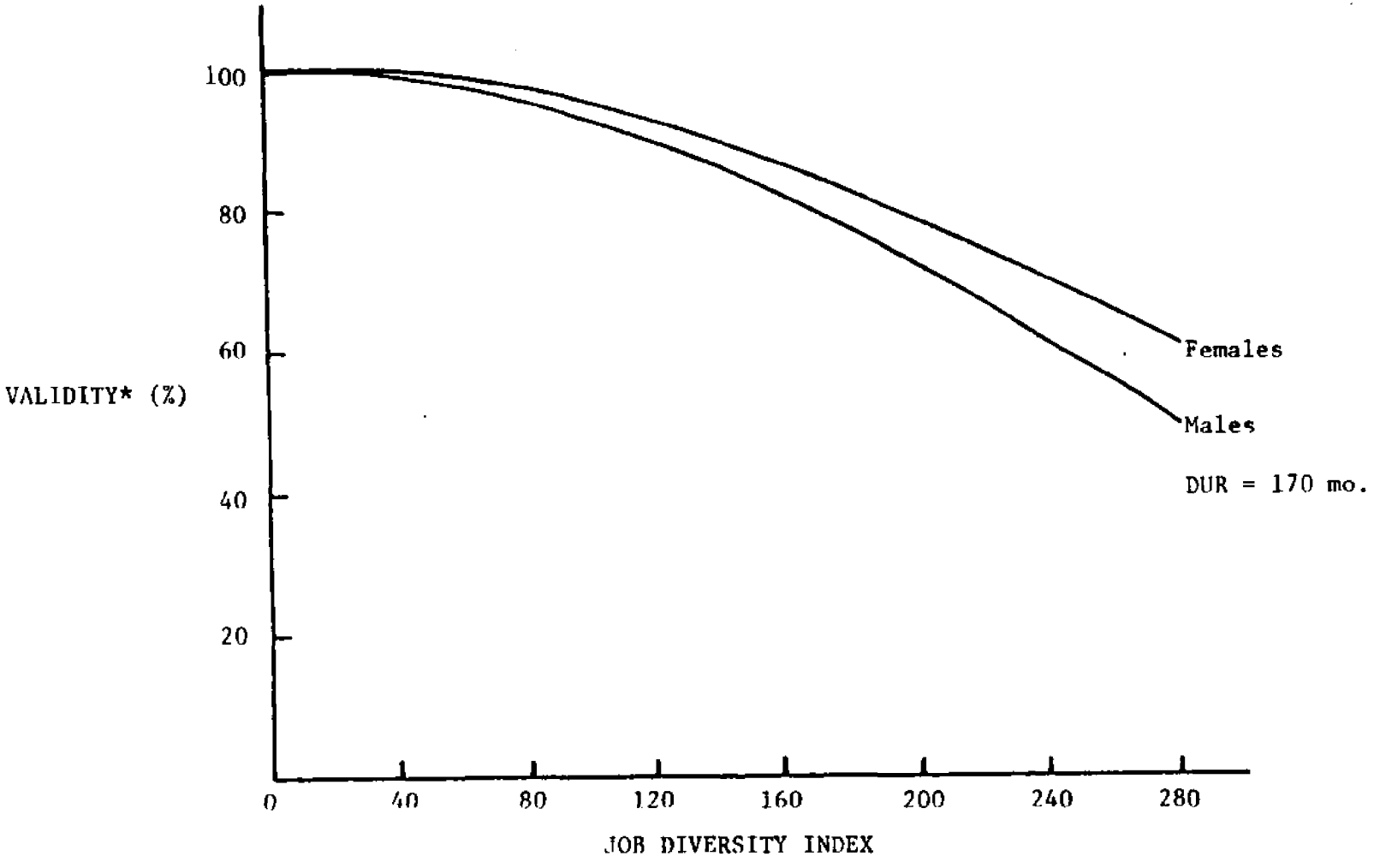


Figure 17. Frequency distributions for crude 1976 validity scores

A. Males

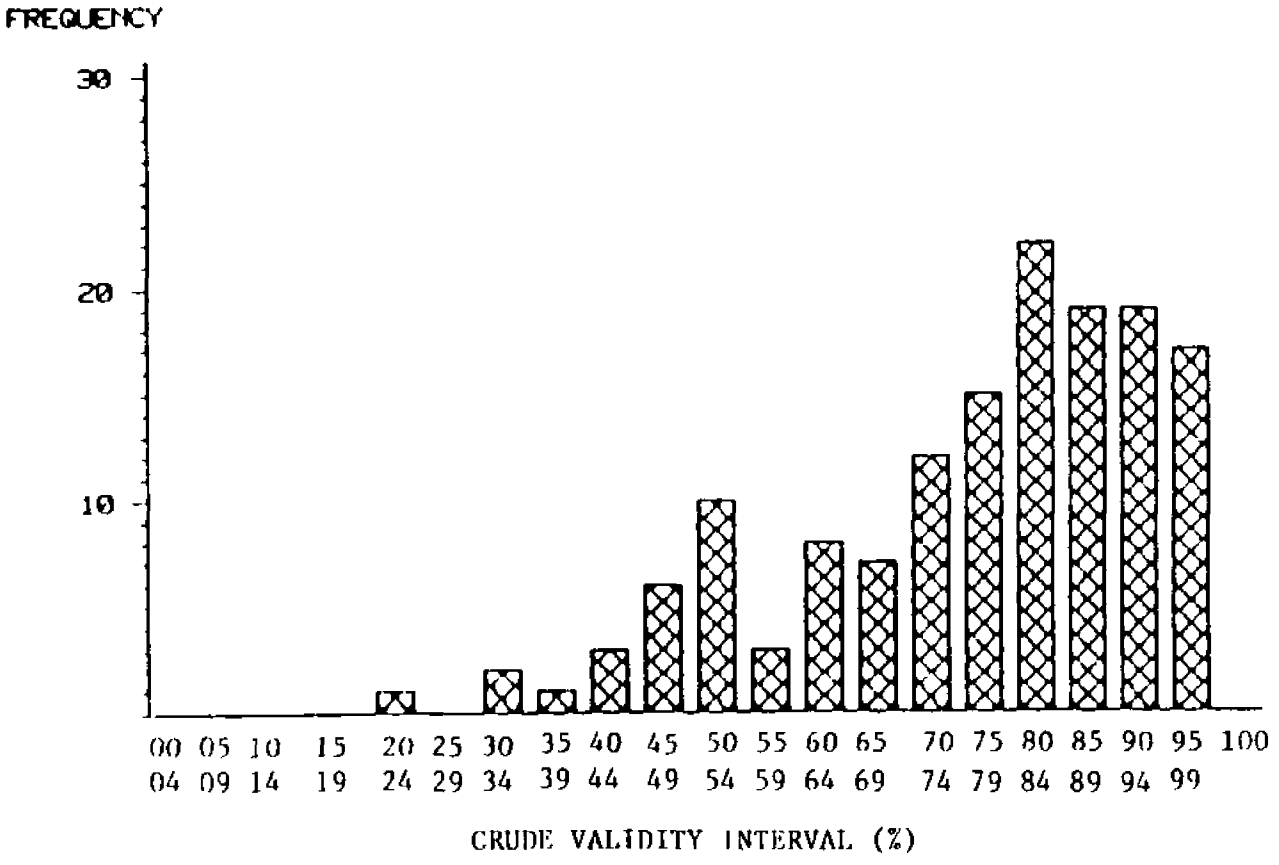


Figure 17. Continued

B. Females

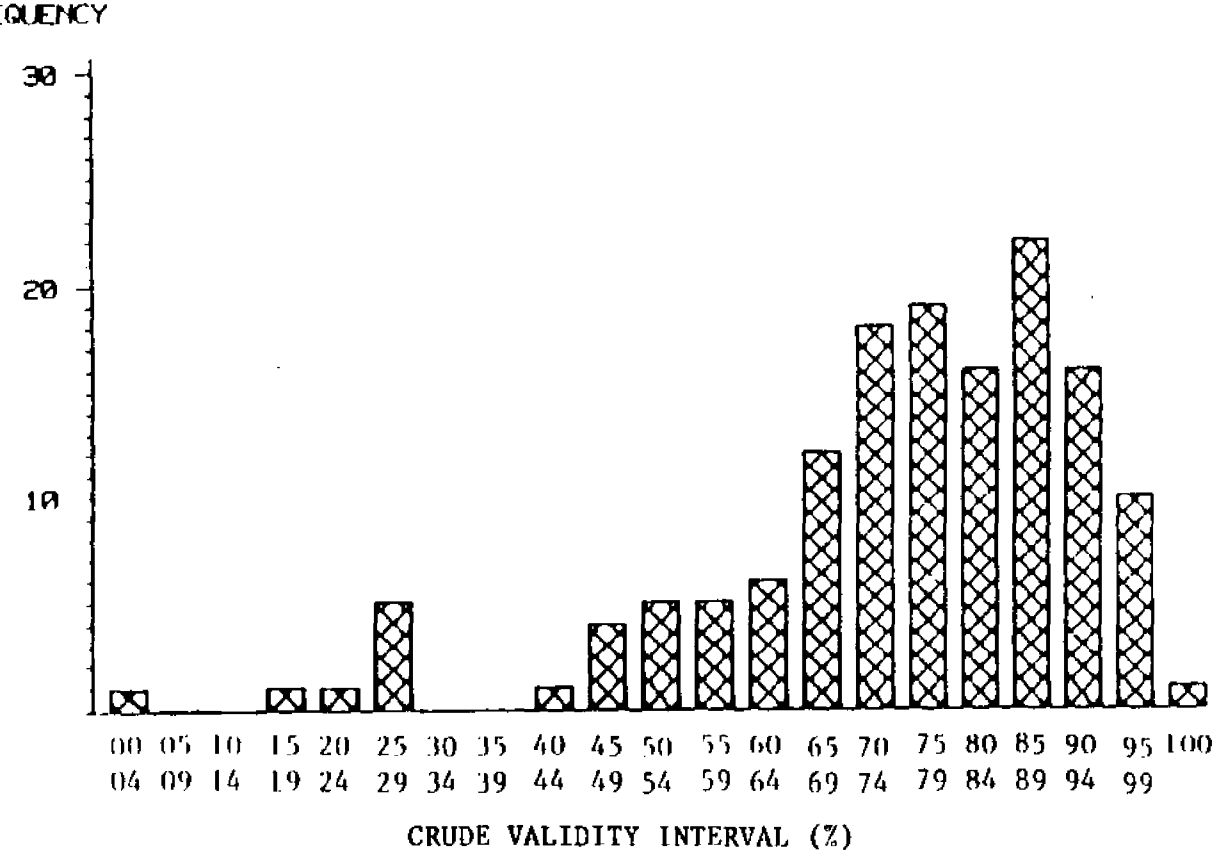


Figure 18. Regression lines describing relationship between validity and examinational delay (reexamined subgroup)

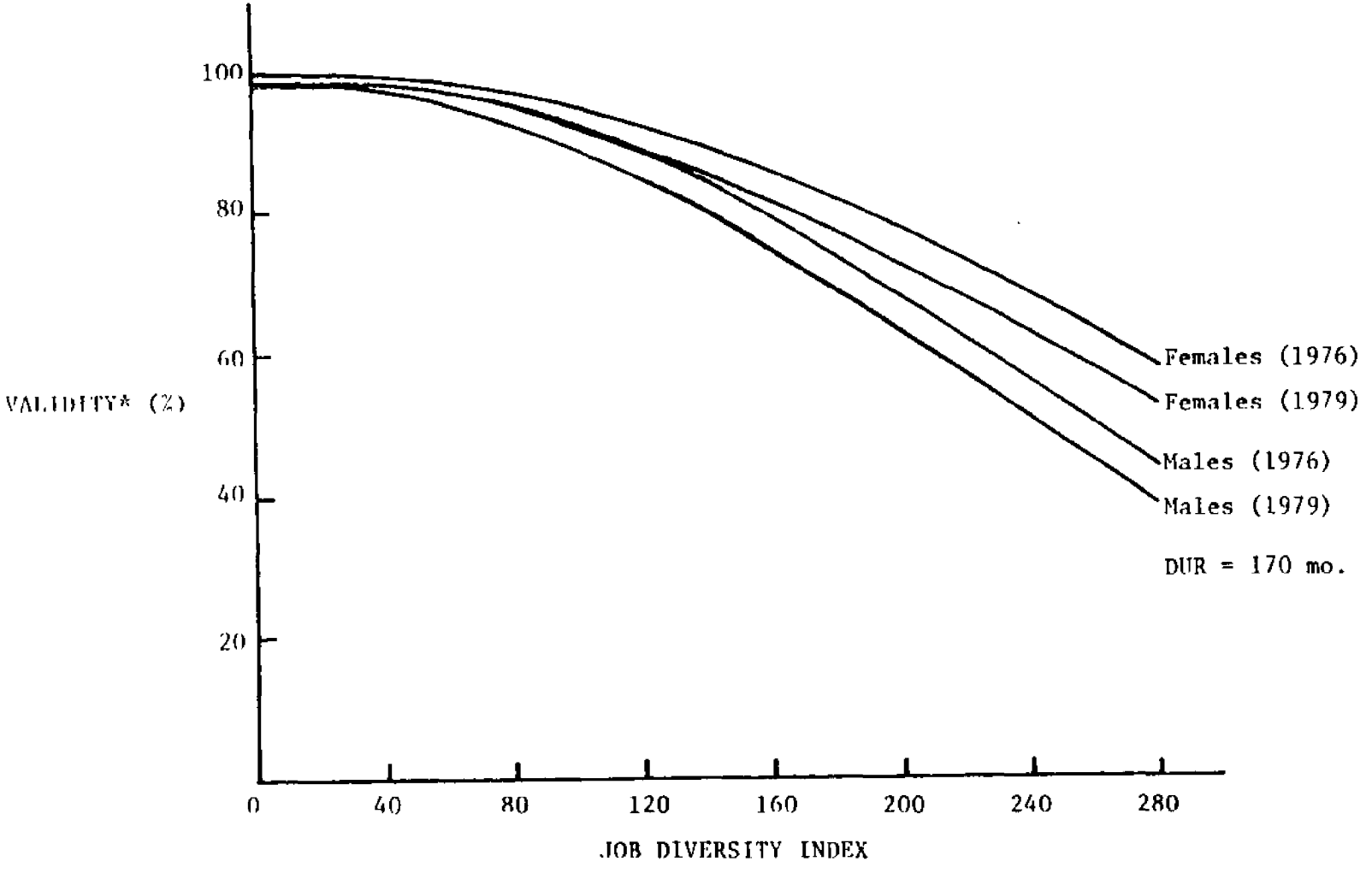


Figure 19. Frequency distributions for crude male 1976 validity scores in reexamined and non-reexamined subgroups

A. Reexamined

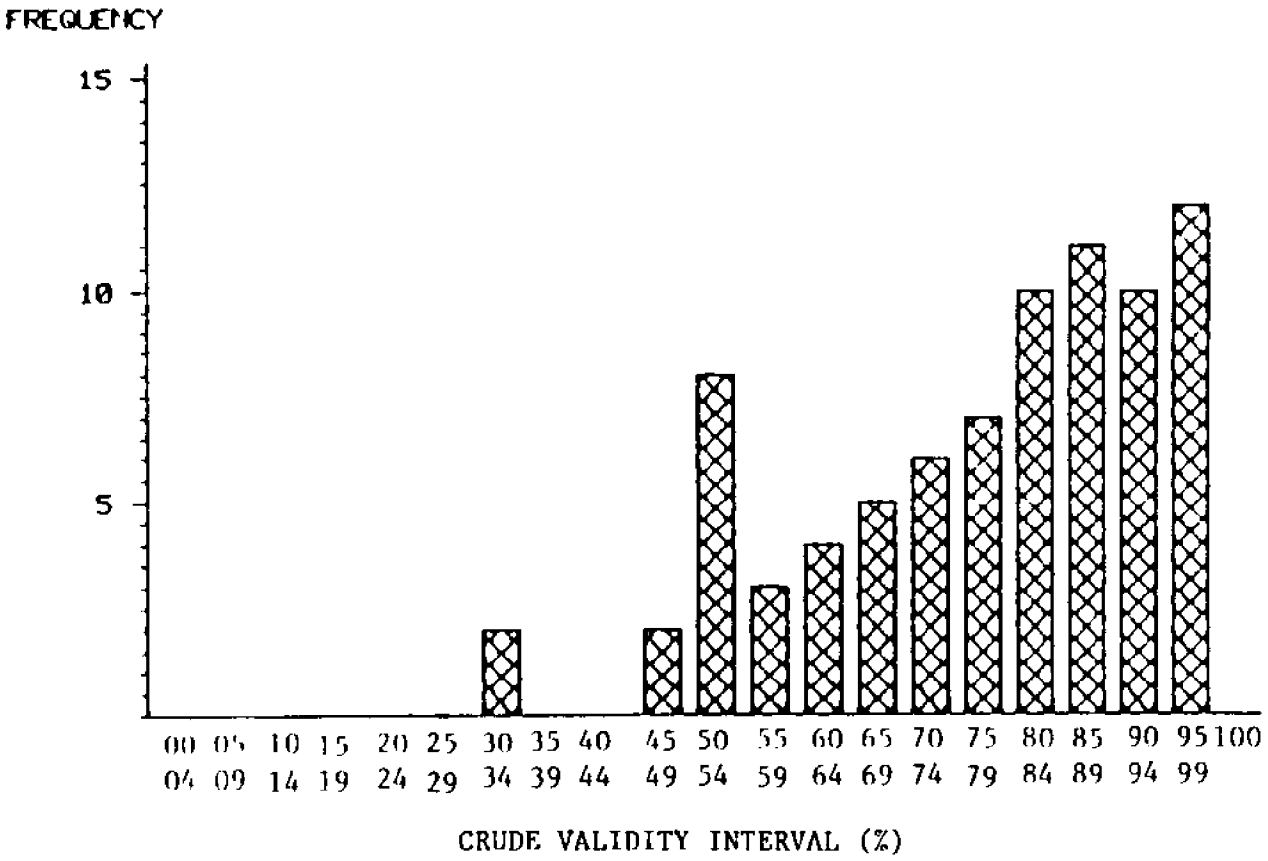


Figure 19. Continued

B. Non-reexamined

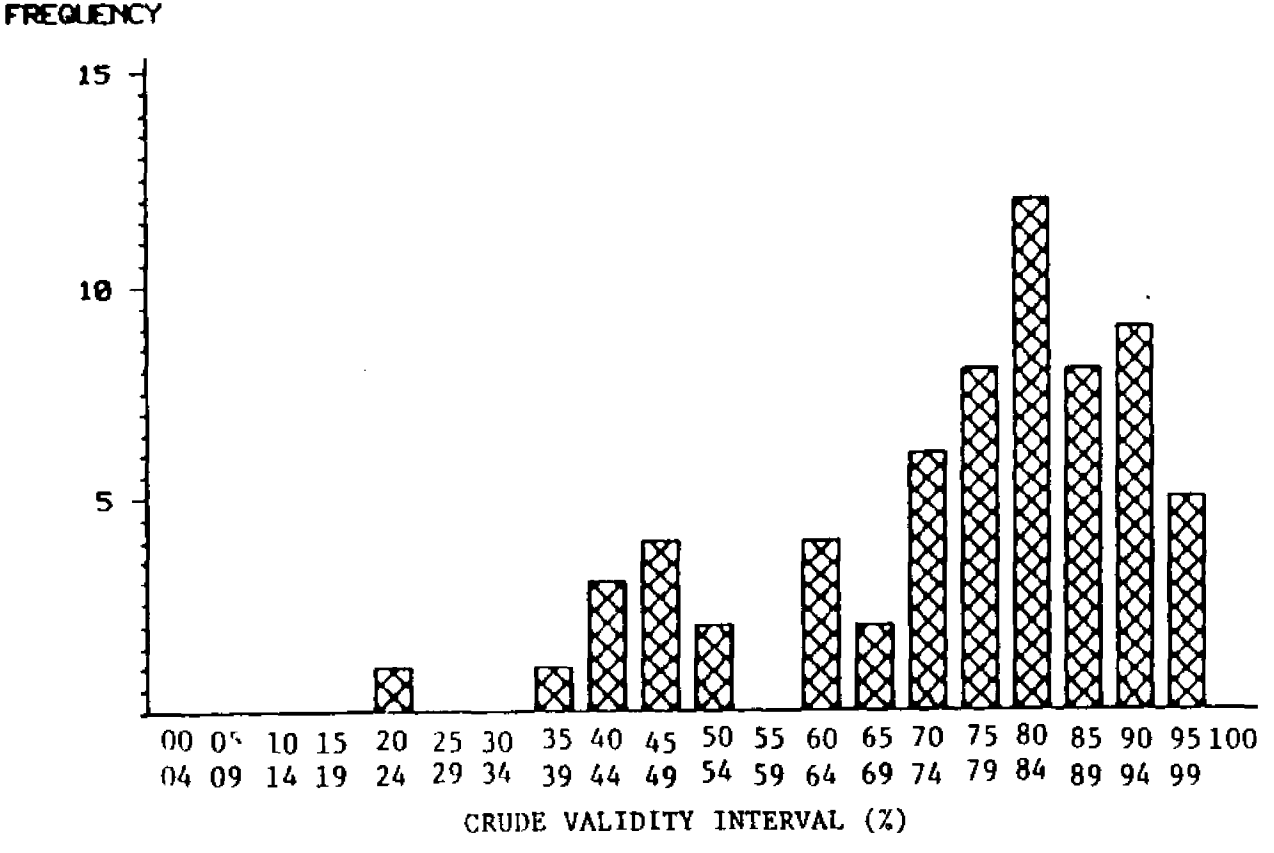


Figure 20. Frequency distributions for crude female 1976 validity scores in reexamined and non-reexamined subgroups

A. Reexamined

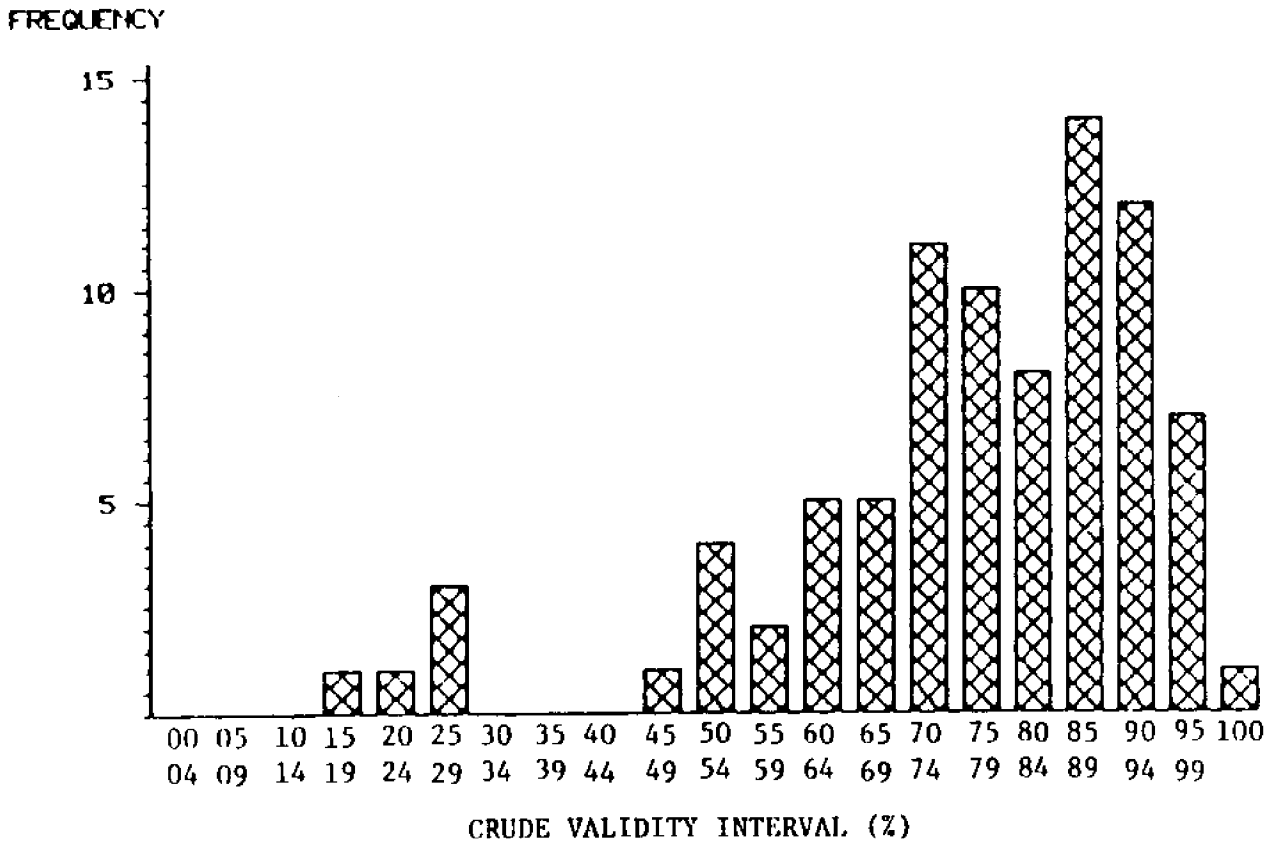


Figure 20. Continued

B. Non-reexamined

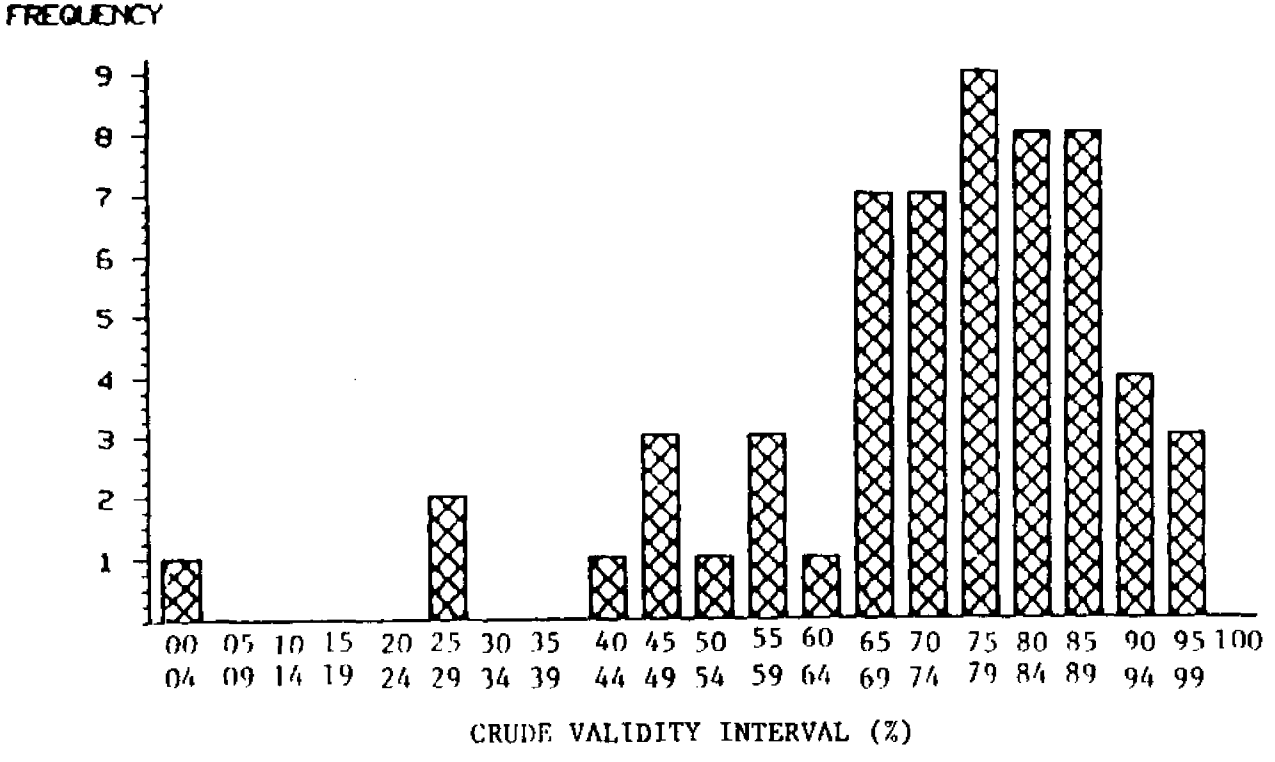


Figure 21. Frequency distributions for crude male validity scores in reexamined subgroup

A. 1976

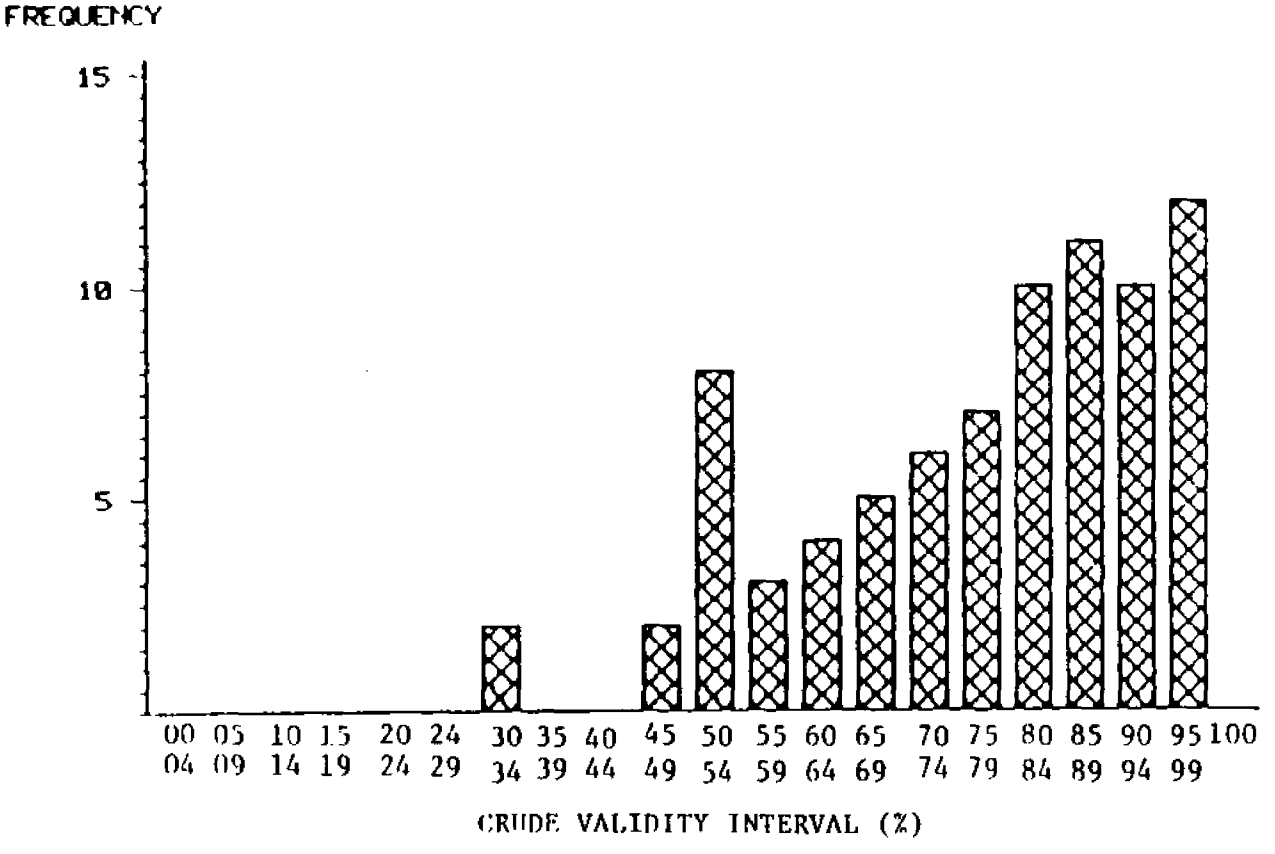


Figure 21. Continued

B. 1979

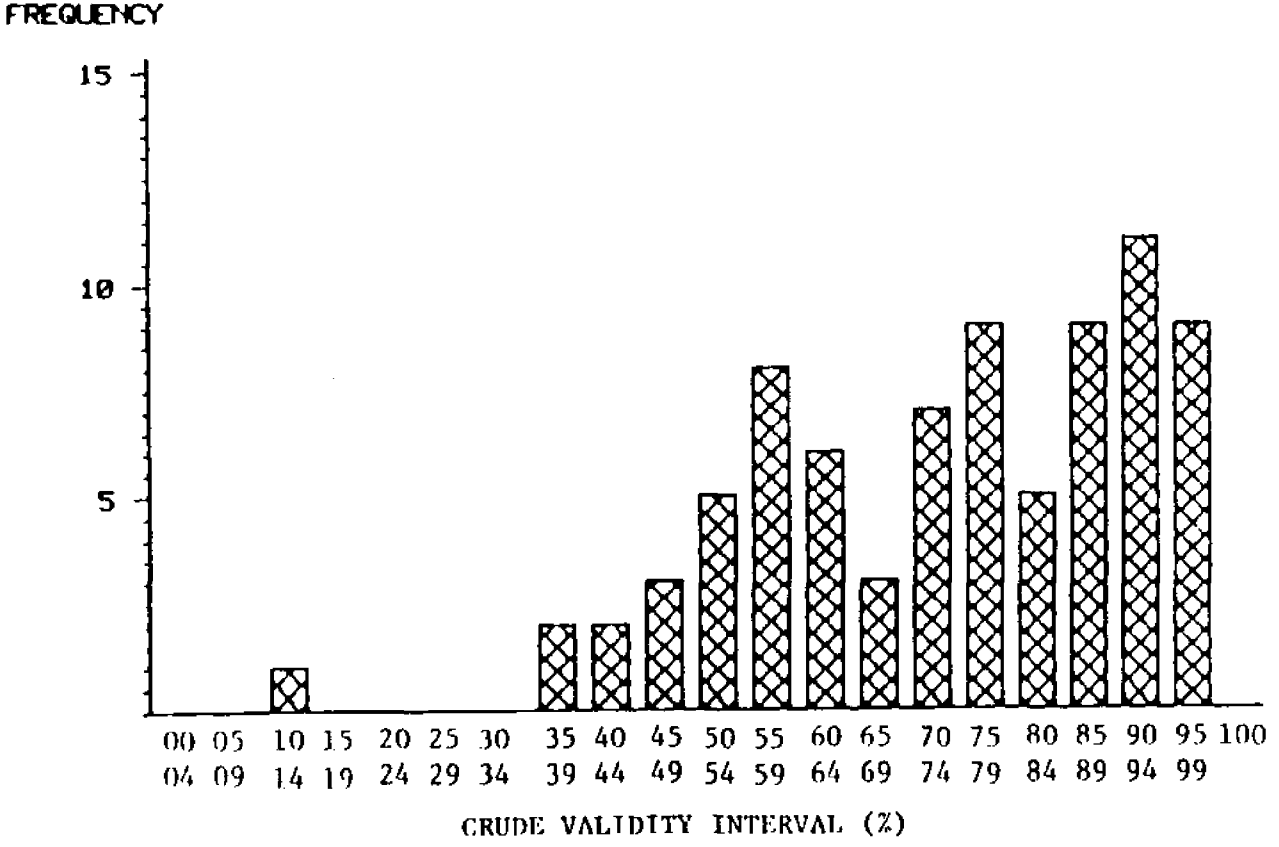


Figure 22. Frequency distributions for crude female validity scores in reexamined subgroup

A. 1976

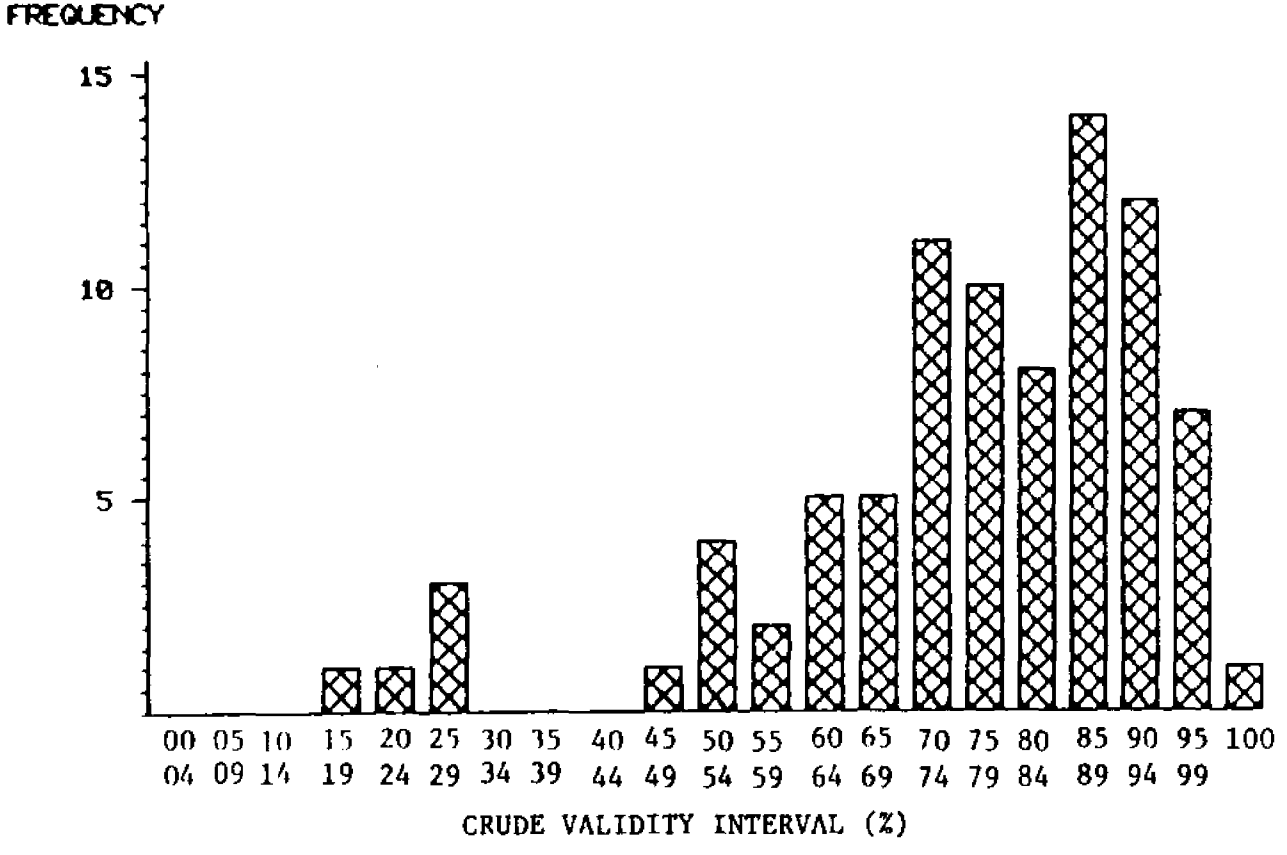


Figure 22. Continued

B. 1979

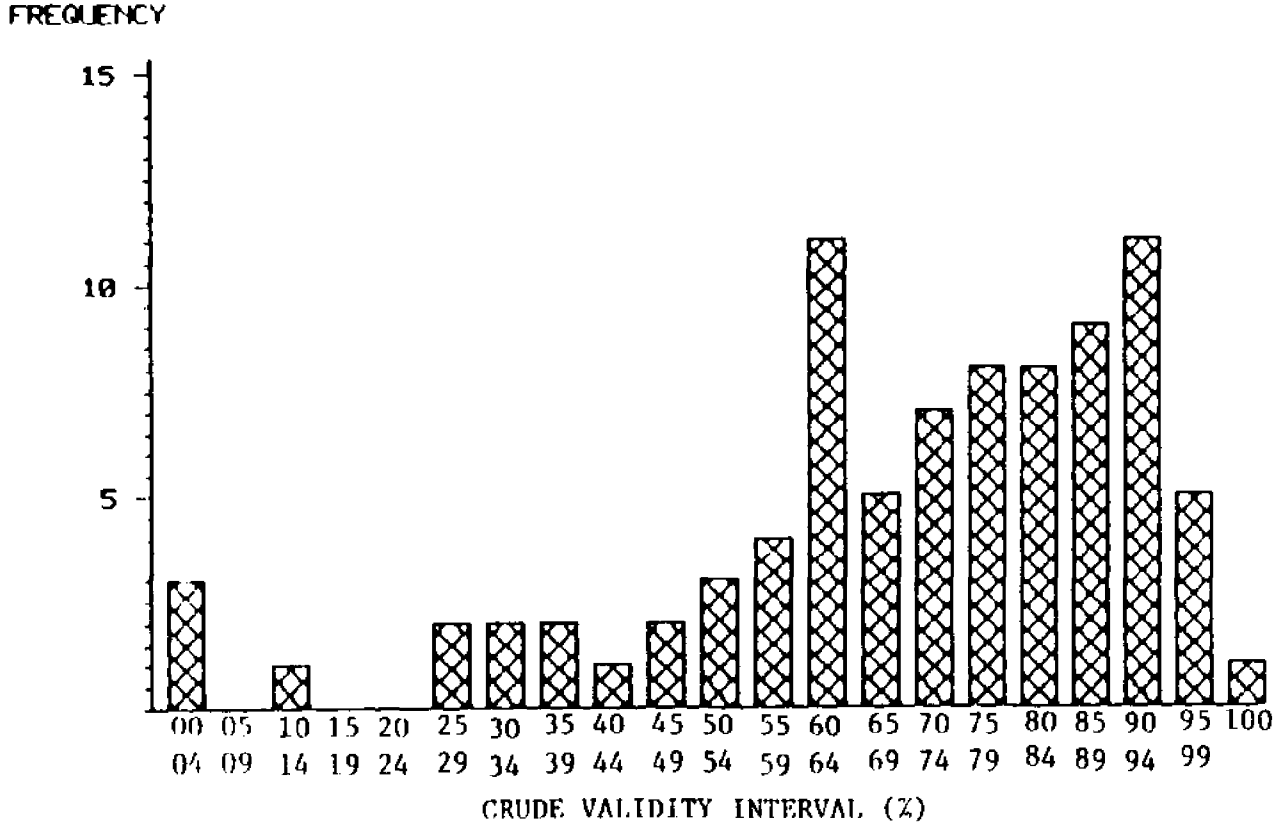


Figure 23. Regression lines describing relationship between 1976 validity and intrinsic time lapse (three subperiod subgroup)

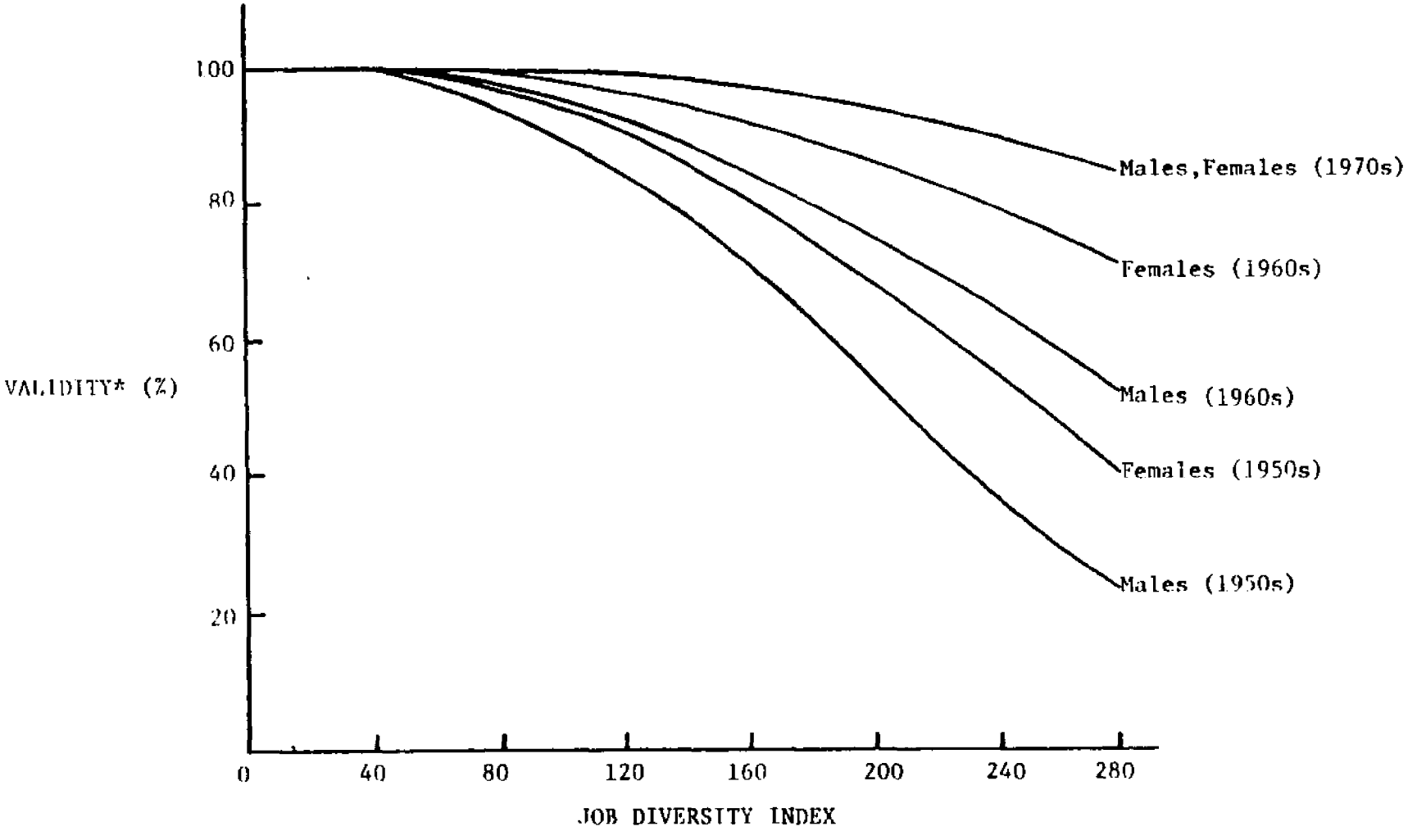


Figure 24. Regression lines describing interviewer effects on 1976 validity

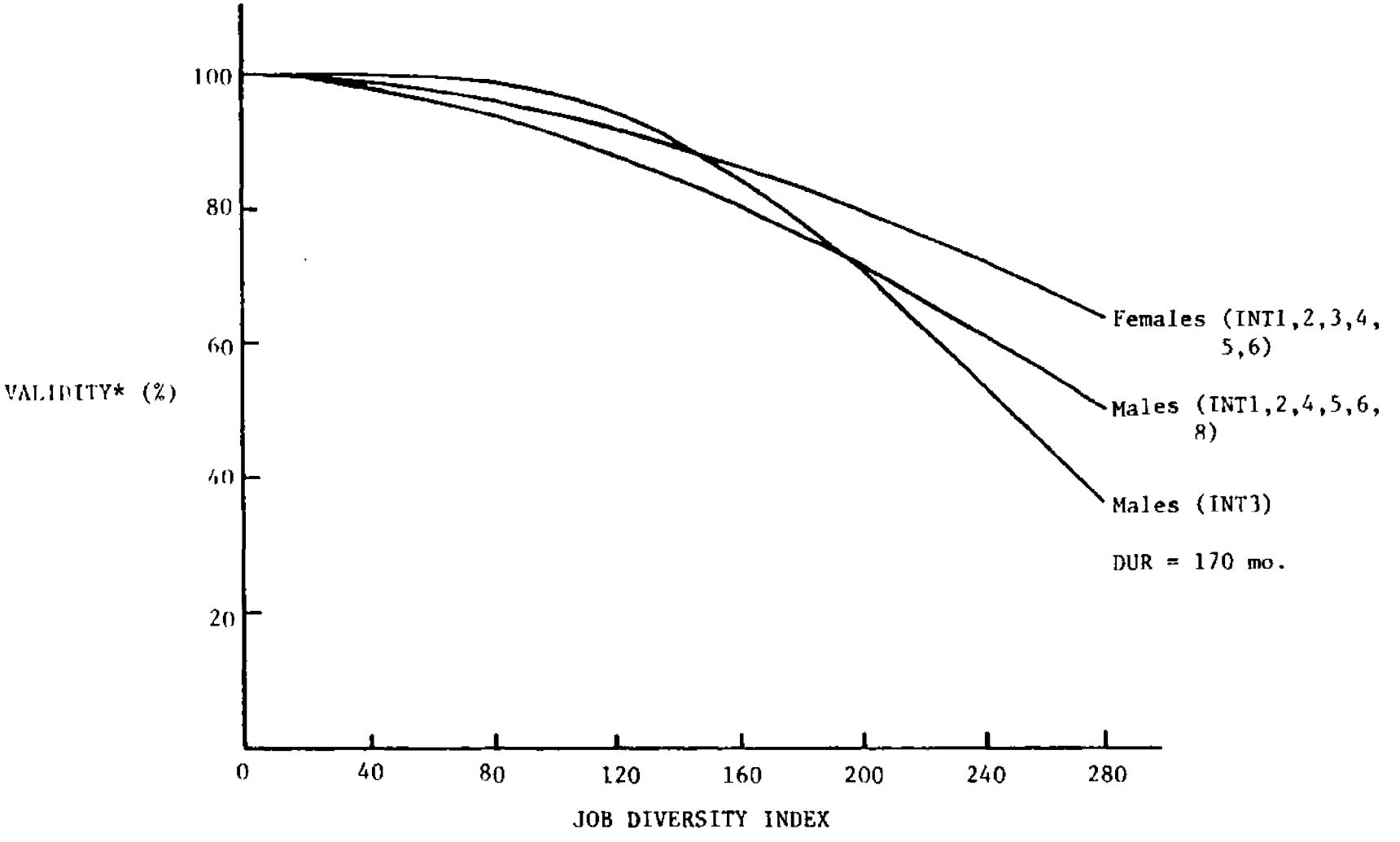
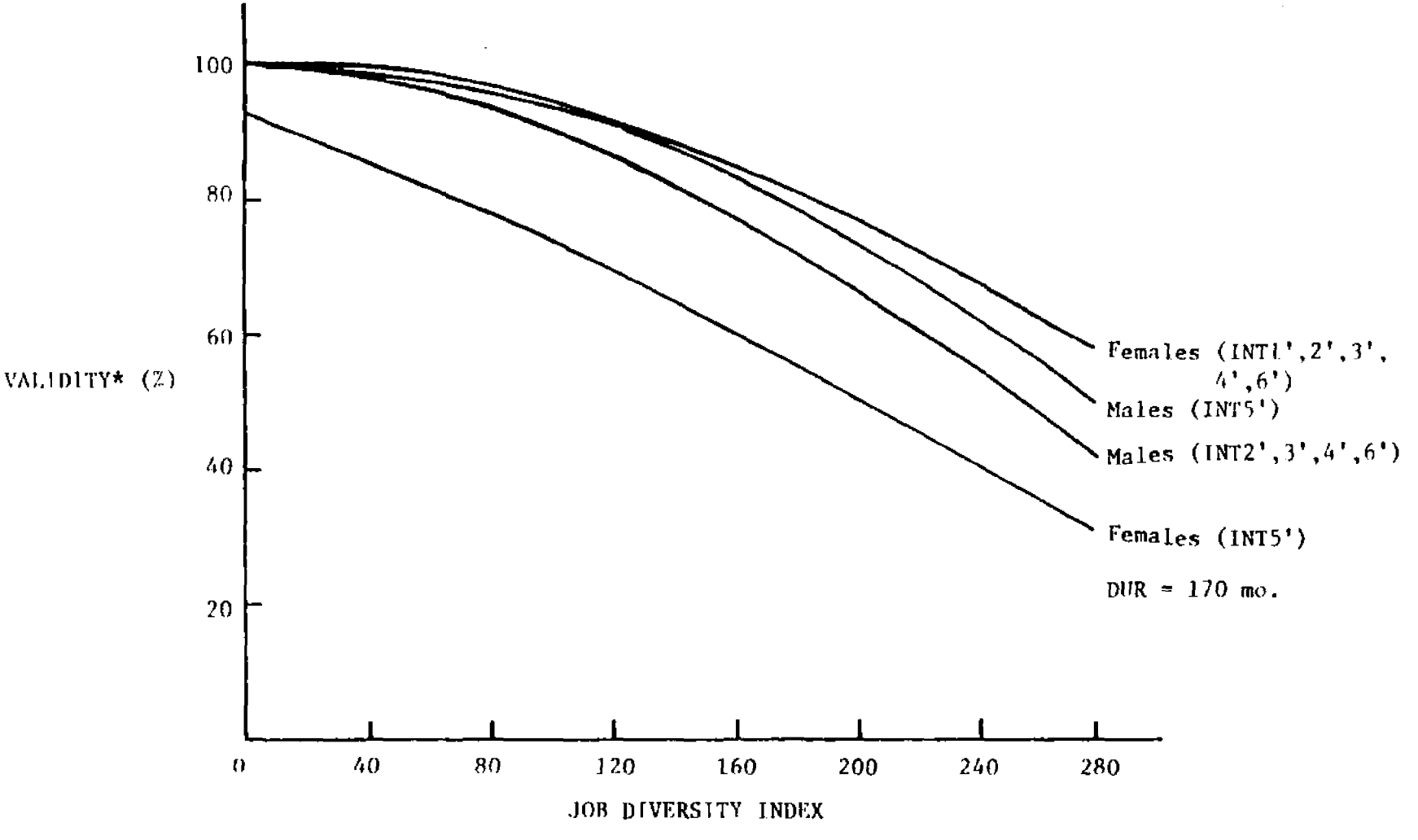


Figure 25. Regression lines describing interviewer effects on 1979 validity



CHAPTER V: DISCUSSION

This study was designed to evaluate the accuracy of self-reported occupational histories obtained from a group of 288 blue-collar workers who were part of an actual survey study of PCB-exposure and effects at two related capacitor-manufacturing plants. Accuracy was measured directly by the operationally-defined parameter of validity. The role of reliability as a possible indirect measure of accuracy will be discussed later. There were numerous factors which could have been responsible for the observed variability in the crude validity scores. In terms of intrinsic study group effects, the causes of inaccuracy were not as important as the fact of their existance since it was this fact that determined actual misclassifications in PCB exposure status. However, if one wanted to apply the validity and/or reliability procedure to evaluate accuracy in other study groups or plan new studies safeguarded against inaccuracies, knowledge of these factors' effects was critical.

These factors were divisible into two classes - directly measureable and intangible. The former group included: the personal attributes of age and sex; the individual occupational factors of duration of employment (DUR) and job diversity index (JDI); the experimentor introduced variables of examinational delay and interviewer; and the temporal factor of intrinsic time lapse. Race, educational level, and socioeconomic status were not listed among the personal attributes since these factors were either homogenous or indistinguishible among the members of this study group. Direct measureability, either on a continuous scale or by class, permitted

analysis of these factors' effects and interactions through the use of multiple regression procedures. Intangible factors included innate recall ability, cooperativity (truthfulness), stress level, and ambiguities in job categorical distinction, etc. While by no means unimportant, these variables could only at best be qualitatively or indirectly estimated and were thus not included in the regression procedure mentioned above. They could perhaps have been symbolically inserted in the regression models as additional "β-terms." Their effects, however, were real and had to be considered when interpreting any of the equations resulting from the regression analyses. One should recall that the regression effects due to measureable factors never did account for all the variability in the validity scores.

To begin the process of sorting out factor effects, the initial multiple regression analysis was performed. This was intended to examine accuracy in the overall sense, with emphasis on the individual personal and occupational variables of JDI, DUR, and sex. Age was eliminated from the analysis on account of the fact that other validity studies (4,6) have shown it not to have a significant negative influence on recall ability. Increasing difficulty of recall has most certainly have been associated with increasing age, but this does not appear to occur to any significant degree in mentally normal individuals of working age. Recall ability may very well erode in normal individuals, but perhaps significantly, only at well-advanced ages. Also partly responsible for an age-recall loss association would be neurological degenerative diseases more frequent with increasing age. However, the overwhelming majority of individuals who participated in this and the cited studies were both mentally sound and not of greatly advanced age. At any rate, age was tied

up to such a considerable degree with DUR, that if it were important, it would have been very difficult to distinguish the two in a regression analysis.

JDI, according to the results, exerted a great influence on validity in this study group. This locally significant effect was indicative of the fact that JDI in the generalized situation would be the ultimate factor or root determinant of accuracy. Intuitively, there could be no variation in degree of accuracy unless the JDI would take on values greater than zero. In this study group, for example, if every worker had remained in a single job category for his or her entire duration (i.e. JDI = 0), there would have been little reason not to expect, all other things being equal, all validity scores to have been at or near 100% (less than 100% only if starting date could not be exactly recalled). In essence, therefore, it takes the introduction of complexity (i.e. room for error) to bring out variation in the degree of accuracy.

Once a level of complexity is defined, the most important influence on degree of accuracy is innate recall ability. This can be thought of as the natural capacity of the evaluated individuals to report the best possible (most correct) history. Other factors under specific circumstances may have a greater influence on accurate self-reporting, but these would tend to vary in importance from study to study or even be absent or insignificant in others. In terms of planning future studies, such effects could possibly be eliminated, minimized, or neutralized. Recall variability, though, will occur in studies no matter what precautions are taken. Investigators must hope that its effect upon accuracy would not be too serious. Recall ability consists of two components: short and long-term memory. The latter has to do with recall of more

distant past events while the former represents a degree of awareness of day to day facts and occurrences.

The JDI and recall ability exert antagonistic effects in determining accuracy. Increasing complexity of the job categorical pattern is a force which limits the overall potential for accuracy. Recall ability attempts to neutralize this force. Figure 26 depicts a proposed mechanism as to how they might interact to determine accuracy in the generalized situation. Pictured is a series of validity distributional patterns corresponding to various JDI ranges. This scheme was based in part on actual results from this study group, i.e. from validity subdistributional frequency histograms resulting from various JDI tercile stratification procedures, the rest on intuitive reasoning.

As stated earlier, it is assumed that at JDI = 0, all validity scores would be virtually 100%. A movement away from the 100% level emerges when JDI starts taking on values greater than zero. When these values are very low, the small amount of complexity introduced is not sufficient to distinguish individuals with varying recall abilities. In essence, those individuals with very low JDI values do not need the aid of good recall ability to give highly accurate reports. Once a threshold level of JDI is reached, and thereafter, the complexity level is high enough to produce potential distributions of validity scores. JDI ranges, i.e. degrees of complexity to be counteracted, define the practical locations and, in effect, limits of these potential distributions, subject, of course, to the effects of local factors. The actual distributions, i.e. degrees to which JDI-defined limits are approached, are determined by the recall ability profile of the particular study group (or subgroup) being investigated. The scheme in Fig. 26 assumes a

random distribution of individual recall abilities in overcoming the various levels of complexity.

Hypothetical study groups (or subgroups) with relatively simple job categorical patterns would produce distributions similar to those on the right side of the figure. The skewed appearance of these distributional patterns, in spite of the assumed random distribution of recall abilities, would be due to the constraint placed upon them by the 100% maximum level for validity. In practical terms, relatively low JDI levels would be unable to distinguish, scorewise, among the upper levels of recall ability. Misclassifications would be no problem for a study group (or subgroup) with the #1 (far right) distribution. They might be more serious for groups possessing the #2 and #3 distributional patterns. Study groups (or subgroups) with more complex job categorical patterns would have potential distributions located more toward the left. With sufficient complexity, the potential distribution would cease to be constrained by the maximum validity level, yielding the normal appearing distribution of validity scores reflective of the assumed random distribution of recall abilities - The entire spectrum of recall abilities would be distinguished (#4). Misclassifications would increase, relative to lower complexity situations, causing serious difficulties. Still higher JDI ranges would produce distributions resembling those on the extreme left of the figure (#6 and #7). Impingement on the minimum validity barrier would account for the skewed nature of these distributional patterns. Practically speaking, the lower recall abilities would be indistinguishible, scorewise. Misclassifications under these circumstances would be overwhelming, making effective studies very difficult, if not impossible.

In this study group, many individuals in a low JDI tercile (JDI: 14-169) were able to score at or near maximum (like the far right distributional pattern). There were still some individuals who were able to overcome the complexity associated with a JDI level of 170-216 and produce very high scores (like the #2 and #3 distributional patterns - females and males, respectively), but they were fewer in number (restricted to those with excellent or very good recall abilities). In a high JDI tercile (217-326), the practical limit of the upper recall abilities was receding from 100% (like the #3 and #4 distributional patterns - females and males, respectively). Recall ability, afterall, is a human characteristic and has its limitations. Therefore, individuals with highly complex job categorical patterns, for the most part, no matter how good their recall abilities, are at a disadvantage in terms of accurate self-reporting. On the other hand, though, the minimum score level is not zero. There probably is no realistic JDI level high enough to produce, by itself, zero validity. Recall ability should be able to overcome at least some part of any degree of complexity. Of course, other factors in specific situations could cause zero validity or aid individuals in producing better than expected scores. Such effects were listed earlier and will be discussed shortly.

Wide ranges of JDI values in a study group would furnish composites of the appropriate patterns in Fig. 26. The wide JDI range in this study group (14-326) produced the composite seen, for example, in the crude 1976 validity score distributions for the original main group (Fig. 17). The ultimate shape of a composite would depend upon the relative numbers of individuals within each JDI subrange.

The 1979 composite male and female distributions (JDI: 14-326, mean crude non-validity of 24-29%) for the reexamined subgroup (Figs. 21B and 22B) were associated with a misclassification rate of 26.8% (HPCB period). Such a rate may have been substantial enough to at least cause a reduced (or even zero) ability to detect any real PCB health effect differences between the exposure categories (if misclassification was nondifferential). At the worst (if misclassification was non-differential), spurious differences could have been introduced. There was, for example, a significantly greater tendency in this subgroup (HPCB period) to overreport amount of time spent in high PCB-exposure job categories. Spurious differences would have resulted if such overestimating occurred to a greater extent in individuals with high exposure (unequal misclassification among exposure categories).

Perhaps the above JDI range-mean crude validity-misclassification rate association could be used as a rough benchmark figure in predicting misclassification rates in other situations with other JDI ranges. For instance, in a study group with a higher JDI range, there would be an expectation of a higher misclassification rate, possibly causing a failure to detect real differences (false negative result). However, this association was subject to various local influences so that any comparisons should thus proceed with caution. The above 26.8% misclassification figure, for example, was quite lenient in that the exposure categories were relatively broad in range and could not be strongly influenced by chronological errors in self-reporting. To clarify, the broad range led to resistance to exposure categorical shifting and thus misclassification. The only chronological factor that could influence exposure category determination was the 1/71 change in PCB mixtures (since it was de-

terminable). Other time factors (i.e. changing plant conditions, etc.) could not be pinpointed and thus could play no role in category placement (Remember that chronological errors were reflected in the crude validity scores.).

The scheme of Fig. 26 was also evident in the validity vs. JDI scattergram (Fig. 14). In this figure, recall ability, among other factors, was represented by the variability about a mean validity line. This variability was quite confined at lower JDI values, but expanded considerably and shifted downward in location as JDI increased.

Sex of the individual, according to the results of the initial regression analysis, also exerted a significant influence on validity scores in this study group. It was an interactive effect, though, dependent upon the level of JDI. The model predicted little or no sex difference in mean validity at low JDI values. However, once sufficient complexity was introduced, the females reported more accurately. The greater the complexity, the wider became the gap between male and female values.

In the context of Fig. 26, increasing JDI caused the progression to the left for each sex at different rates. In a low JDI tercile for the original main group (14-169) both sexes possessed distributions like #1. Low levels of complexity made any advantage conferred by being female inconsequential (i.e. The sex effect threshold was not reached.). In a 170-216 JDI tercile, whatever accounted for the female advantage was able to distinguish them from their male counterparts in terms of accuracy. These females had a distribution like #2 while that for the males was more like #3 (hinted earlier). For a high JDI tercile (217-326), females and males possessed distributions like #3 and #4, respec-

tively. Returning to the regression model, the threshold for the sex effect was at JDI = 40. Somewhere above JDI = 280, although not shown by the model, the male regression line should start to level off as a minimum validity level was approached.

Whether this female advantage was real, and thus had practical implications for the generalized situation, was difficult to say with certainty. It was highly statistically significant. Some other studies of accuracy have found women to report better than men (5,9). Possible explanations could lie in male-female differences in several of the listed intangible factors. One of these, innate recall ability, was just discussed in conjunction with the JDI. Two other factors which may have been responsible for a sex difference were cooperativity and stress level. Of course, the degree to which sex was tied up with levels of these or other pertinent factors would determine how universal this effect was.

Cooperativity was reflected in a subject's attitude, interest in the study, and ability to communicate with the interviewers. Intuitively, it is a factor that would tend to increase accuracy. The fact that individuals were willing to participate in a study in the first place selected for a relatively higher degree of cooperativity than found in the target population. However, there was still room for variability, subsequent to selection. In two studies of validity, cooperativity, as estimated by subject responses to a question concerning their willingness to allow the experimentors to release data of a personal nature, was found to vary (4,6). Furthermore, the fact that there usually are certain subjects in almost every study who feel they have something to gain by influencing a certain outcome and proceed to distort their re-

porting (i.e. truthfulness), is additional evidence for cooperativity variation. The same studies which showed this cooperativity variability also reported that those subjects who responded to the particular question in a positive manner did indeed have a higher group validity level than those who did not.

High stress level, either on the part of the subject or intensified by the conditions imposed by the examiners would most certainly have had a negative influence on accurate reporting. Stress levels have been indirectly estimated through various physiological responses but were not directly measureable in this study situation.

In terms of sex differences concerning these intangible factors, females have been known to handle stress better than males. During an evaluation proceeding, lowered stress level could mean less interference with the recall process. Females may also be more cooperative and/or less distracted than males in self-reporting situations. This could be expressed in terms of better rapport with interviewers; a greater interest in or concern with the study; a more positive, submitting, or accommodating attitude toward the examiners; or a lower probability of distorting information intentionally. As to recall ability, perhaps females could have either a natural or acquired (out of necessity) advantage. In numerous reporting situations concerning family disease or exposure history, females have been the preferred source of information. During such procedures they seem to have a better grasp of time and dates (i.e. dates of births, pregnancies, illnesses, etc. - markers of time). An additional consideration was that interviewer preference toward female subjects could have caused the sex difference in validity. The implications of this will be discussed later.

In this study group, males and females had crude mean validity scores that were not significantly different, seemingly contradicting what was shown by the initial regression procedure. Based upon such results, one could have assumed a similar situation in other study groups (or certainly no female advantage). Of course, this wasn't true - The explanation lied in the fact that female JDI levels in this study group were significantly higher than those for males. The former were thus put at a significant local self-reporting disadvantage which at least compensated for their regression-predicted advantage. In other words, the male-female JDI difference confounded an underlying sex effect (possibly generalized) in this study group (This confounding was corrected for by JDI adjustment in the regression procedure.).

Also highly significant in the initial regression analysis was DUR. It was quite interesting that its effect upon accuracy was positive. This reinforced the arguement that age was not (within the context of what was discussed earlier) a significant negative influence on accuracy. The fact that age and DUR were tied together to such a great degree and produced a positve effect argued against age being a strong negative influence.

The explanation for the positive effect of DUR on accuracy may lie in its association with other factors. At first glance, intuitively, DUR would seem to be a negative influence on accuracy in the sense that a longer work history would be relatively more difficult to recall. However, a longer history would only be more significantly difficult to recall, as will be discussed in detail later when considering intrinsic time lapse, if the history were sufficiently complex (high enough JDI). As to positive influences, remember that DUR was a net measure (absolute

duration minus lost work time). Therefore, especially for females (who tended to lose much work time), increased DUR meant fewer work interruptions (layoffs, illnesses, etc.) and a more stable, less complex work history - a positive influence on accuracy. This cannot fully explain the positive effect, though, since DUR also significantly increased male accuracy. Recall that males did not tend to lose much work time. It could be that DUR reflected intangible factor differences between long- and shorttime workers (see Chapter III), the former having higher accuracy perhaps due to greater identification and familiarity with the plant operations or higher cooperativity. At any rate, possible negative intangible effects on DUR were outweighed by the positive ones. Some of the above DUR effects were undoubtedly tied up with JDI (through influencing complexity), but JDI and DUR correlation was only minimally positive.

In terms of Fig. 26, increasing DUR, given a constant range of JDI, would favor a progression toward the distributional patterns on the right. Obviously, if JDI were very low (i.e. distributional pattern on the far right) any assistance of high DUR would be superfluous in terms of accuracy. Analagously, in Fig. 16, the slopes of the regression lines would become less negative. The significance of the DUR effect in the generalized situation might tend to be variable. Nevertheless, it might perhaps be closely tied to the typical blue-collar setting.

The next factor of interest was one which would be introduced by the experimentor - examinational delay. Specifically, this represented the elapsed time from the end of the work history period to the time of examination. In this study group, the work history period ended as of 3/76 - the time of the original examination. Subgroup reexamination was

done 45 months later, in 12/79. This latter examination compared with this same subgroup's original one measured the effect of 45 months examinational delay. Delayed examination intuitively would cause lower accuracy than would have occurred with immediate examination. The longer the period between the occurrence of events and their documentation the less is the liklihood of accurate recall.

Examinational delay was handled by the second multiple regression analysis. Separate sets of regression lines were generated for 1976 and 1979 subgroup validity. Like the sex effect, examinational delay was not a disadvantage to accurate reporting when complexity of the job categorical pattern was low. At a threshold JDI level of about 40, the same as for sex, the 1976 and 1979 mean value validity lines began to diverge (i.e. sufficient complexity introduced). Unlike with sex, though, the mean value lines kept a constant difference of 5 to 6 points after JDI passed the 160 level. This indicated an intercept rather than an interactive effect. However, counteracting examinational delay in this study group to some extent was a "learning effect." That is - being evaluated a second time might have aided recall. Such aid may have been more important for those respondents who had more complex work histories (i.e. in an interactive way - those with more complex patterns had more opportunity to be aided by the learning effect). Without the confounding of the learning effect, examinational delay may very well have produced an interactive effect (i.e. a more negative slope for later examination). Note again that the sex effect was intact for each examinational period.

In Fig. 26, the effect of examinational delay would be a shift toward the leftward distributional patterns (except in a group with the far right distribution - little or no shift in this case). Adding on

the sex effect, this shift would be greater for males. Incidentally, JDI tercile stratification on the 1979 subgroup validity scores produced actual distributions up to #5.

The effect of examinational delay represents a difficulty with which any investigator must contend when making use of self-reported information. The longer the delay, the greater would be the effect upon the recall of events (possibly interactive minus the learning effect - increasing 1976-1979 gap with higher JDI). The obvious and only way to minimize its effect would be to gather information concerning exposure as close in time as is possible to the actual period of occurrence.

In this study group (specifically the reexamined subgroup) a 45 month examinational delay was reflected by a significant 4 (for males) to 7 (for females) point drop in mean crude validity scores (underestimate due to "learning effect"). This local delay effect represented an overall composite figure - the outcome of many individual delay effects which varied with JDI range (i.e. low JDI - low delay effect, medium and high JDI - 4 to 7 point delay effect).

Reexamination, which permitted the analysis of examinational delay, also allowed an accuracy comparison between reexamined and dropout individuals. Intuitively, a number of those who dropped out may have done so on account of lack of interest or cooperativity. If this number was substantial and the above were true it could have been reflected in lower dropout accuracy. There was a local decrease of 2 to 4 points (not significant) from reexamined to dropout individuals. Whether or not this represented a significant association could not be established here (would need a larger sample size or more homgenous comparison), but it would be worth future investigation.

Intrinsic time lapse (function of the individual) brought out the fact that the degree of accuracy was not a constant over the entire recall period, but rather a composite of numerous subperiod accuracy figures. Intuitively, the farther back in time the subperiod, the lower would be the chance for accurate recall. A third multiple regression procedure was used to handle the intrinsic time lapse effect. Just like the sex factor and perhaps EXP without the learning effect, intrinsic time lapse was a function of JDI. A threshold level of complexity was necessary to bring it out. Logically speaking, a worker who stayed in a single job category for his or her entire duration would have little trouble accurately reporting for the most recent, intermediate, or earliest subperiods.

The complexity introduced by JDI values greater than 120 was sufficient to produce the expected decade trend. The effect was intensified (i.e. greater accuracy difference between subperiods) with further JDI increases. As to the sex effect, it remained intact only in the earliest two subperiods. This meant that the female advantage was more complex than indicated in the overall analysis. It was interactive with both JDI complexity and intrinsic time lapse. In other words, low complexity or less than five years recall was insufficient to distinguish male-female accuracy differences. Also present at high JDI and earliest subperiod was a levelling off of male scores as basement validity levels were approached. A phenomenon counteracting intrinsic time lapse effect was "first job recall." The initial category in an individual's working career was a significant event (a marked occurrence in time) and was thus recalled more easily. Notable was the fact that the JDI effect was somewhat dependent upon intrinsic time lapse. Scores for the most re-

cent subperiod (1970-76) were somewhat resistant to JDI effects whereas the earlier decade subperiods were significantly affected (mutual dependence of decade and JDI effects). In essence, JDI is a powerful influence on accuracy, but not all powerful.

The further back in time the subperiod, given sufficiently high JDI, the more the validity distributions would resemble those on the left side of Fig. 26. Some of the left side distributional patterns of this figure were actually observed in this study group (for the 1950-59 subperiod).

Intrinsic time lapse effect would have important implications for any self-reporting retrospective epidemiological study. Such studies looking back upon five years of exposure history would present fewer problems than one requiring ten or more years of recall.

In terms of local effects in the three subperiod subgroup, the intrinsic time lapse effect was quite evident. No local confounding conditions were able to minimize it, much less conceal it. Concerning misclassifications - If one had been specifically interested in PCB-exposure during the 1950s subperiod, the rate of misclassification would have been much higher than any of the overall figures (17.2 to 26.8%).

To the factors examined in the initial regression analysis (JDI, DUR, and sex) was added an additional one, that of interviewer (INT). To handle this enlarged set of factors, a fourth regression analysis was performed. The INT factor was excluded from the original analysis since overall effects and individual variables were of specific interest. INT was a study-specific, examiner-introduced, controlled, and correctable variable. Certain self-reporting studies have been conducted without the assistance of interviewers. The reasons for their exclusion varied

from financial (they are an added expense) to attempts to reduce outside bias (i.e. Interviewers have been known to introduce errors that outweigh their data collecting advantage.). When employed, their number makes a big difference. In small studies with a single interviewer there could be a positive, negative, or neutral influence on accuracy, but relatively less variance (no subjective inter-interviewer differences). With two or more interviewers, the effects on accuracy can vary more. In this study, there were 16 interviewers (9 in 1976 and 7 in 1979). This many were necessary on account of the large number of subjects to be processed within a short period of time. The question was if differences among interviewers in the skills necessary to obtain the optimum self-reported history: patience, perseverence, objectivity, intuition, and ability to communicate with the subjects, could yield significantly different interviewer-specific accuracy profiles.

According to the regression analysis results, significant interviewer-specific validity differences did exist. These differences were both of the both main (location of regression line) and interactive (slope of regression line) types. Main effects in the generalized sense would most likely reflect interviewer differences in technique, skill, or objectivity - independent of extraneous factors and static over the long run and across different study situations. In this study, the main effect was due to an objectivity difference. INT5' in 1979 demonstrated a definite sex preference, producing at most JDI levels constant, significantly higher male and lower female validity scores relative to those of other interviewers. A skill or technique difference would have been reflected by a non-specific intercept effect. Slope effects, on the other hand, would most likely reflect interviewer differences in re-

sponse to, for example, difficulty of cases (reflected by JDI). In this study, INT3 in 1976 had a significantly more negative interaction with JDI relative to the other interviewers. Curiously, this effect was restricted to male subjects. In addition, INT3 had a significant main effect preference for male respondents. The net result for this interviewer was relatively higher male scores at low JDI levels (easy cases) but increasingly lower male scores above JDI = 200 (difficult cases). The interactive effect, in practical terms, more than compensated for the intercept effect.

It should be noted that even though only two interviewers produced significant validity differences, there was considerable variation of other interviewer mean values about the composite regressopm line of Fig. 16. Undoubtedly, with larger sample sizes per interviewer and a more homogenous group, a greater number of significant variations would have been confirmed.

As to the possibility of interviewer preference accounting for the overall female advantage, this seemed unlikely. The fact that three of the four significant effects either favored males or disadvantaged females (with females still keeping their overall advantage) mitigated against this possibility. In addition, the highly negative interactive effect of INT3 on male validity was not sufficient to account for the overall male disadvantage, i.e. removing the INT3 slope effect still left a significant male disadvantage.

INT5' (females 1979) pointed out the fact that interviewers could greatly compound accuracy difficulties or even negate the advantage of low JDI. Conversely, highly trained interviewing specialists may be able to help subjects overcome large JDI disadvantages. While

such effects could not be certified as generalizable, in reality they may be present in varying degrees in many study situations (Of course, they would tend to be unpredictable.). In the scheme of Fig. 26, positive interviewer effects favored distributional patterns on the right and vice versa for negative effects.

Interviewer error (i.e. negative effects upon accuracy) in this study group was expressed in lack of date recording, incomplete job descriptions, inaccurate coding, poor organization, and overlooked information. In terms of the latter, recall that regular work time was lost for various reasons. Such losses were rarely picked up during the interview unless they were for greatly extended periods (i.e. greater than six months). It was evident that the interviewers stressed the reporting of positive events (i.e. When and where did you work?) while excluding the negative ones (i.e. When did you not work?). A four-month long strike (10/68 - 2/69), for example, did not show up on but a single self-reported history. The fact that females tended to lose more regular work time than males meant that this type of interviewer error put them at an additional local disadvantage in terms of accuracy (producing a further underestimate in this study of an actual female advantage).

One miscellaneous factor which affected accuracy in this study group was ambiguities in job categorical distinction. The effect of this factor was negative and resulted from both occupational and examiner-introduced conditions. Certain job categories may have been difficult for the workers to define or distinguish, leading to variable entries on the occupational history form. This may have been exacerbated by the categorical scheme provided by the examiners. Interviewers increased the chances for this type of error by sometimes failing to

record job descriptions along with the titles.

In summary, the acquisition of a non-recall influenced source of exposure information (company employment records) subsequent to the conductance of a particular retrospective prevalence study (the surveys of Fischbein et al.) which depended upon self-reported information to determine exposure status, made it possible (in a new study - the present one) to objectively test the accuracy of such self-reported information, including a determination of rates of misclassification resulting from any inaccuracies. This was achieved through use of the validity parameter. Then, statistical manipulations of the local validity scores, using multiple regression analysis, provided an opportunity to gain insight into the conceptual (hypothetical, generalized) value of the self-reported work history as an instrument in reasonably representing individual exposure profiles. In essence, the validity analysis here permitted a spot check of the accuracy of a study already performed to be utilized: 1) by Fischbein et al. in assessing the efficacy of their surveys, 2) in the establishment of guidelines toward preventing or minimizing inaccuracies in the planning and execution of future self-reporting studies, and 3) in making decisions as to whether or not a particular study should be undertaken at all.

An analysis of validity, while being useful in terms of performing the latter two functions, would be of little practical value as an accuracy-testing tool in specific self-reporting study situations. Investigators undertaking self-reporting studies would logically not expect to obtain objective sources of exposure information. If they could get such information it would not make much sense to employ self-reported information and worry about accompanying inaccuracies and misclassi-

fications. They would use the objective information. Self-reported information in such cases could still be useful, but only in providing supplementary data that may have been missed in using the objective sources alone (Refer to the Chapter II concerning the shortcoming of the company employment records.).

In terms of accuracy evaluation in specific self-reporting situations, the parameter of reliability would enter the picture. A reliability procedure would involve a repetition of the work history, testing for consistency (matching of intervals) between the two versions (no validity base required). In the strictest sense, the reliability parameter would not be measuring non-validity - the difference between the work history and the actual history (systematic error - a methodological problem involving study design), but rather imprecision. Precision error reflects random flucuations (variance) of the self-reported work history in how it estimates the actual work history (attributable to sampling variation). Intuitively, though, there is reason to believe, in terms of self-reporting, that these two error types might cover some common ground. That is, a reliable self-reporter might very well be a valid reporter, or better yet, an unreliable reporter would most certainly be an invalid reporter. If so, reliability (non-reliability) may be useful in some study situations in providing an indirect indication of validity (non-validity). However, one must exercise caution prior to invoking this procedure - There would always be some degree of uncertainty associated with it: There could be an excess of consistent, yet inaccurate self-reporters in a particular study group or a reasonably accurate original work history could be followed by a low quality second history, the latter resulting from poor interviewing technique.

Timing (i.e. gap period) could also be a problem for a reliability analysis. The investigator must consider the time delay between the original and necessary (for reliability) reexamination. A set of examinations too close in time would allow the subjects' initial reporting to greatly influence their repeat reporting. This would not give a true test of reliability (9). A sufficiently long delay that would allow subjects to "forget" or not be influenced greatly by their original reporting would yield an unbiased reliability figure. Perhaps with a longer delay, such as the 45-month gap in this study, (i.e. with ample time for long-term memory to be affected) reliability may become a better predictor of validity. Consider, in this study, the initial 1976 history being a substitute validity base which the subjects upon reexamination would attempt to reproduce in a manner analagous to his or her attempt to recall the company records during the initial evaluation (i.e. similar process yields validity-reliability relationship). This substitute validity base, however, would be less complex than the original (easier to reproduce) probably yielding reliability scores overestimating validity. However, this may be balanced by the fact that reliability, being based upon two examinations and interviewers (validity was only based upon one), would allow a greater chance for error. Unfortunately, the conductance of such a procedure was not valid in the context of this study. Any calculated reliability figures (1976 history vs. 1979 history) would have had to have used the same denominator, recall period, as that used for validity, such that a relationship would automatically have been imposed between the two parameters. In other words, it would have been impossible to determine if any observed validity-reliability relationship were real or spurious.

CONCLUSIONS AND RECOMMENDATIONS

Logically, when existent, objective sources of exposure information should be employed in conducting retrospective epidemiologic studies. Increasingly, employers and unions, with the cooperation of government and medical personnel, are making this possible by beginning to or being in the process of keeping track of employees' work histories in industries where possibly harmful occupational exposures may occur. Even so, there will continue to be cases where objective sources will not be available, or if obtainable, not for various reasons, satisfactory (i.e. incompleteness, lack of necessary specificity, subjective alterations or recordings, illegibility, delayed availability, appearance of new or unexpected diseases with unknown or unrecorded exposure patterns, etc.). The only alternative in such cases would be to use self-reported information to assess exposure status. The question is - Is the self-reported retrospective history satisfactory enough?

The validity analysis in this study demonstrated the local and conceptual implications of using retrospective self-reported information. Concerning the former, particular rates of misclassification were contrasted with specific degrees of accuracy. This information was of interest to Fischbein et al. who have been and still are in the process of relating outcome data (health effects) to PCB-exposure levels. Depending upon the ultimate aims of their study (i.e. what they hoped to establish) they will have to decide whether or not the misclassification rates found here could have seriously affected their study (How much could calculated prevalence odds ratios have differered from the true values?). In terms of future studies, the validity analysis pointed out both potential problems and certain ground rules. The ground rules will

always be established by the factors of JDI and innate recall ability. The JDI, as depicted in the scheme of Fig. 26, will set the realistic (predictable) potential limits on accuracy. The actual distribution should reflect the random spectrum of human recall abilities - Recall ability attempts to neutralize or counteract JDI. In this context (ignoring factor effect levels), it should be pointed out that inaccuracy caused by high JDI level is study-specific and that an individual presenting an inaccurate work history in a particular study for that reason would not necessarily do so in another situation. On the other hand, an invalid individual due to poor recall ability would carry this attribute to any study situation. All other factors can either attenuate or increase the above basic factor effects as described earlier. Note that the JDI effect was analagous to that of extrinsic factors influencing validity in several papers analyzed in the literature review (e.g. unknown conditions and subclinical infections).

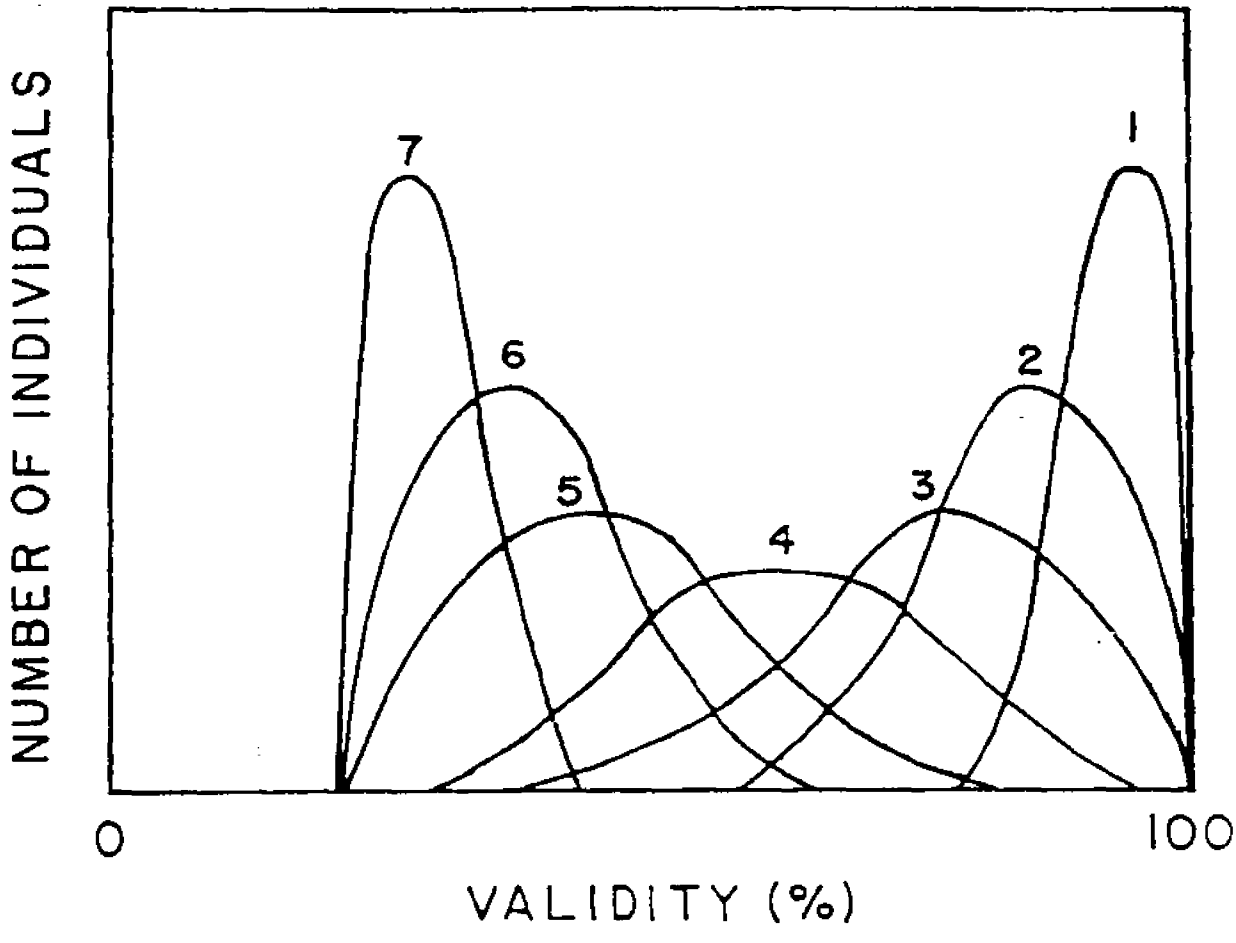
Knowledge of all factor effects and their implications by future investigators could make for reasonably accurate self-reporting studies (assuming that they should have been performed in the first place). The obvious first step would be to conduct the study as close in time as is possible to the occurrence of the exposure events - understanding the detrimental effects of examinational delay and intrinsic time lapse. The second important step involves the reporting protocol: 1) Subjects should be instructed to prepare (in writing if possible) for the reporting session in advance. 2) The environment should be conducive to encouraging optimum reporting (low stress, high cooperativity, etc.). 3) Highly trained interviewers should be employed if necessary and in optimum numbers (minimum of overwork and inter-interviewer differences).

Among other things, they should be trained in memory-prodding techniques and in eliciting both positive and negative events (e.g. job categories and lost work time). 4) Occupational history forms and the interviewing process should be structured and unambiguous. Finally, the male reporting disadvantage should be taken into consideration.

A reliability analysis could possibly be of use in evaluating accuracy in specific study situations. This evaluation, of course, would be subject to the various limitations discussed earlier and must be timed properly. Under some circumstances it could give a rough estimate of study efficacy. At the very least it would measure precision error which after validity is an important consideration.

As a test of the methodology in this study, further research is recommended on other study groups with various JDI ranges, factor levels, exposures, or population types when the opportunity arises. Confirmation of the JDI effect and the reliability-validity relationship would be especially important. Finally, a situation where the exposure agent under investigation is a memory-affecting neurotoxin should be considered. This would add an interesting piece of complexity to any validity analysis.

Figure 26. Validity distributional patterns predicted by JDI level



LITERATURE CITED

1. Dunn, J.E., Jr. and Buell, P. 1959. Association of cervical cancer with circumcision of sexual partner. J. Nat. Cancer Inst. 108: 749-64.

2. Lerman, S.J., Lerman, L.M., Nankervis, G.A., and Gold, E. 1971. Accuracy of rubella history. Ann. Intern. Med. 74(1): 97-8.

3. Goebel, W.M. 1979. Reliability of the medical history in identifying patients likely to place dentists at an increased hepatitis risk. J. Am. Dent. Assoc. 98(6): 907-13.

4. Madow, W.G. 1967. Interview Data on Chronic Conditions Compared with Information Derived from Medical Records. National Center for Health Statistics Report Ser. 2, No. 23, U.S. DHEW. Public Health Service, Washington, D.C.

5. Commission on Chronic Illness. 1957. Chronic Illness in the United States. Vol. IV. Chronic Illness in a Large City: The Baltimore Study. Harvard University Press, Cambridge, Massachusetts.

6. U.S. National Health Survey. 1961. Health Interview Responses Compared with Medical Records. Health Statistics Series D, No. 5. U.S. DHEW. Public Health Service, Washington, D.C.

7. Petitti, D.B., Friedman, G.D., and Kahn, W. 1981. Accuracy of information on smoking habits provided on self-administered research questionnaires. Amer. J. Pub. Health 71(3): 308-11.

8. Brady, W.F. and Martinoff, J.T. 1980. Validity of health history data collected from dental patients and patient perception of health status. J. Am. Dent. Assoc. 101(4): 642-45.

9. Sacks, J.J., Krushat, W.M., and Newman, J. 1980. Reliability of the health hazard appraisal. Am. J. Pub. Health. 70(7): 730-32.

10. Corwin, R.G., Krober, M., and Roth, H.P. 1971. Patients' accuracy in reporting their past medical history: A study of 90 patients with peptic ulcer. J. Chronic. Dis. 23: 875-79.

11. Chamberlain, G. and Johnstone, F.D. 1975. Reliability of the history. Lancet 11 Jan 75: 103.

12. Norell, S.E. 1981. Accuracy of patient interviews and estimates by clinical staff in determining medication compliance. Soc. Sci. Med. 15E(1): 57-61.

13. Andrasik, F. and Holroyd, K.A. 1980. Reliability and concurrent validity of headache questionnaire data. Headache 20(1): 44-6.

14. Streissguth, A.P., Martin, D.C., and Buffington, V.E. 1976. Test-retest reliability of three scales derived from a quantity-frequency-variabilty assessment of self-reported alcohol consumption. Ann. N.Y. Acad. Sci. 273: 458-66.

15. Thompson, J.K. and Collins, F.L., Jr. 1979. Reliability of headache questionnaire data. Headache 19(2): 97-101.

16. Sanders, B.S. 1962. Have morbidity surveys been oversold? Amer. J. Pub. Health 52: 1648-59.

17. Paganini-Hill, A. and Ross, R.K. 1982. Reliability of recall of drug usage and other health-related information. Am. J. Epidemiology 116(1): 114-22.

18. Meltzer, J.W. and Hochstim, J.R. 1970. Reliability and validity of survey data on physical health. Public Health Rep. 85(12): 1075-86.

19. Pecoraro, R.E., Inui, T.S., Chen, M.S., Plorde, D.K. and Heller, J.L. 1979. Validity and reliability of a self-administered health history questionnaire. Pub. Health Rep. 94(3): 231-38.

20. Schlesselman, J.J. 1982. Case-Control Studies: Design, Conduct, Analysis. Oxford, New York.

21. Fischbein, A., Wolff, M.S., Lilis, R., Thornton, J., and Selikoff, I.J. 1979. Clinical findings among PCB-exposed capacitor-manufacturing workers. Ann. N.Y. Acad. Sci. 320: 703-15.

22. Jones, M. 1978. Industrial Hygiene Survey of the Two Capacitor-Manufacturing Plants. Industrial Hygiene Section, Industrywide Studies Branch, Division of Surveillance, Hazard Evaluations and Field Studies, NIOSH, Cincinnati, Ohio.

23. Kleinbaum, D.G., Kupper, L.L., and Morgenstern, H. 1982. Epidemiologic Research: Principles and Quantitative Methods. Lifetime Learning Publications, Belmont, California.

24. MacMahon, B. and Pugh, T.F. 1970. Epidemiology: Principles and Methods. Little, Brown and Company, Boston, Massachusetts.