# CANADIAN THESES

# THÈSES CANADIENNES

## NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

## AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

NL 339 (r. 86/01)

Canada

96



NAME OF AUTHOR/*NOM DE L'AUTEUR* George Tien

TITLE OF THESIS/*TITRE DE LA THÈSE* Offense Seriousness: A Test of the Power Model and Interaction

UNIVERSITY/*UNIVERSITÉ* Simon Fraser University

DEGREE FOR WHICH THESIS WAS PRESENTED/ *GRADE POUR LEQUEL CETTE THÈSE FUT PRÉSENTÉE* Doctor of Philosophy (Psychology)

YEAR THIS DEGREE CONFERRED/*ANNÉE D'OBTENTION DE CE GRADE* 1983

NAME OF SUPERVISOR/*NOM DU DIRECTEUR DE THÈSE* Ronald M. Roesch



DATED/*DATÉ* Aug. 12, 1983 SIGNED/*SIGNÉ* [redacted]

PERMANENT ADDRESS/*RÉSIDENCE FIXE* [redacted]

# OFFENSE SERIOUSNESS: A TEST OF THE POWER MODEL AND INTERACTION

by

George Tien

B.S., University of South Carolina, (1969)

M.A., University of South Carolina, (1971)

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY (PSYCHOLOGY)

in the Department

of

Psychology



SIMON FRASER UNIVERSITY

August 1983

APPROVAL

Name: George Tien

Degree: Doctor of Philosophy

Title of thesis: Offense Seriousness: A Test of the Power Model and Interaction

Examining Committee:

Chairperson: Dr. Bernard Lyman

Dr. Ronald Roesch

Dr. Raymond Koopman

Dr. Raymond Corrado

Dr. Vincent Sacco
Alternate Member

Dr. Stephen Golding
Associate Professor
University of Illinois
External Examiner

Date Approved: August 11, 1983

PARTIAL COPYRIGHT LICENSE

I hereby grant to Simon Fraser University the right to lend my thesis, project or extended essay (the title of which is shown below) to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users. I further agree that permission for multiple copying of this work for scholarly purposes may be granted by me or the Dean of Graduate Studies. It is understood that copying or publication of this work for financial gain shall not be allowed without my written permission.

Title of Thesis/Project/Extended Essay

Offense Seriousness: A Test of the Power Model and Interaction.

Author:

(signature)

George Tien

(name)

Aug. 12, 1983

(date)

## ABSTRACT

A currently popular method for indexing the seriousness of offenses was developed by Sellin and Wolfgang. The seriousness index is based on Stevens' power model. The original research and many subsequent studies have focussed on the construction and the use of such an index. Little attention has been given to the technical aspects of scaling. A number of these issues are therefore examined in the present study.

Whether or not the power model fits the subjective offence seriousness data has not been explicitly studied. In this study, a goodness of fit test was developed to assess the fit of the power model.

Magnitude estimation procedures produce seriousness scales with arbitrary units of measurement. The units are determined by the personal preference of the subject or that of the experimenter. Researchers have often not faced up to the fact that comparisons of scale values based on different units of measurements are inappropriate. A method for overcoming the unit of measurement problem was developed for use in this study.

Investigators have long been at odds with one another regarding the additivity of offence seriousness ratings. With the unit of measurement problem resolved, a more definitive test for additivity was now possible.

The results of the study indicated that male and female subjects differed in their perceptions of offence seriousness. Nevertheless, the general pattern of the results for the two

groups was very similar. For both males and females, the power model was found not to be appropriate. Whether this lack of fit is real or due to procedural artifacts cannot be at present determined, because of a general lack of comparative data. A number of transformations were attempted. Neither the threshold transform nor the log-limit transform produced any improvements. However, for both sexes, the quasi-linear transform produced a substantially improved fit. The data also indicated that Sellin and Wolfgang were correct in assuming additivity of crime types, as no interaction was found in both the male and female groups.

## DEDICATION

To my wife, Amy, and to my parents.

## ACKNOWLEDGMENTS

This work is a reflection of the combined talents of many individuals and I would like to thank those who have contributed to the these and substance of this thesis. First, I must acknowledge Dr. Ronald Roesch for his guidance and his encouragement throughout the development of the thesis. I would like to thank Dr. Raymond Koopman for his invaluable input and knowledge which was responsible for the rigor with which the study was designed and carried out. I would also like to thank Dr. Raymond Corrado for his guidance and comments on the various drafts of the thesis. I must also thank Alison Hatch for her assistance in data collection. Her technical expertise in the use of Textform was also greatly appreciated. Finally, I must thank my in-laws, Mr. and Mrs. Phillip Wong for their patience and their constant support.

TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# I. Introduction

A currently popular method for indexing the seriousness of offenses was developed by Sellin and Wolfgang (1964). The classification system was developed in response to the dissatisfaction expressed by many researchers in regard to the inability of the Uniform Crime Reports, a crime index published by the Federal Bureau of Investigation, to reflect the qualitative aspects of a crime.

## Classification Based on Broad Legal Classes

One frequent criticism of the Uniform Crime Reports is the charge that the broad legal classes of the Reports tend to mask the real nature of the offence. For instance, classifying a crime as just a robbery tends to mask the specific nature of the crime. Sellin and Wolfgang (1963) for example, remarked that a robbery could be the armed holdup of one or more persons, resulting in injury to one or more victims, and the theft of thousands of dollars, or it could be the taking by a youngster of a few pennies from a younger child under threat of a beating. According to the Uniform Crime Reports, a robbery is a robbery, regardless of the degree of injury to victims or amount of property loss, as long as no one is killed or raped in the

event.

## Compression of Multiple Offenses

If a victim is killed during a robbery, the event is no longer classified as a robbery, but instead is classified as a murder or non-negligent manslaughter. The records would not show that a robbery has also taken place. Only the highest offence is reported in the Uniform Crime Reports.

## Attempted Acts

Under the Uniform Crime Reporting system, attempted acts are mixed with completed acts. For example, no distinction is made between an attempted rape and an actual rape. Both acts are recorded as forcible rape.

Given these deficiencies, the development of a crime classification procedure which takes into account the qualitative nature of a crime by Sellin and Wolfgang (1964) was therefore considered by some (Wellford & Wiatrowski, 1975; Kelly & Winslow, 1970) to have marked an important advance in the study of criminology.

Sellin and Wolfgang (1964) were primarily concerned with the construction and use of a seriousness index. They were less concerned with the theoretical bases of the term seriousness as

it applied to criminal offenses. They did not operationalize their concept of seriousness. It is conceivable that different subjects focussed on different dimensions of "seriousness" such as moral outrage or social and personal retribution. However, in spite of this lack of clarity regarding the meaning of seriousness, most researchers are willing to assume that perceived seriousness is unidimensional and is well understood by respondents so that no explicit definition is necessary. The issue of the dimensionality of perceived seriousness of offenses needs to be addressed in a rigorous manner. This study, however, focuses on the more technical aspects of scaling, and their implications on the resulting seriousness index.

## II. Literature Review

The Sellin and Wolfgang (1964) crime classification system resulted from a series of studies which involved both interval scaling and magnitude estimations. Their aim was to develop an index for delinquency by making use of Stevens' power law. The power law relates subjective magnitude to stimulus magnitude. The form of this relationship, according to Stevens (1957), can be expressed as a power function. That is, sensation is a power function of stimulus magnitude.

Seventeen raters took part in an initial trial rating of 141 offenses, in which a 7-point category scale was used. Following the initial rating, the median score for each offense was calculated. Three items were then selected to represent each of the seven scale divisions. In all, a total of 21 items were selected. The remaining 120 offense descriptions were randomly grouped into four sets of 30 descriptions. In a second round of ratings, four groups of subjects were used. These consisted of a group of 38 juvenile court judges, a group of 286 police officers and two groups of students, 163 from Temple University, and 82 from the Ogontz Penn State Center. Approximately half of the members of each group, 279 subjects in all, rated the seriousness of the offenses using a magnitude estimation procedure, and the remaining subjects rated the seriousness of the offenses using an 11-point category scale. Raters with the

magnitude estimation task were given a "standard" offense with a seriousness score of 10 and were instructed to rate the other offenses relative to that standard. Each subject rated the set of 21 selected offenses, and some members of each group had the additional task of rating one of the four sets of 30 other offenses. Thus, unlike the trial run, no single rater rated more than 51 offenses. A set of 20 items was eventually selected and was presented to 195 students from the University of Pennsylvania. In this final round of ratings, a magnitude estimation procedure was used. Their choice of magnitude estimations over interval ratings is not surprising, since one of their purposes was to determine if Stevens' power law could be applied to "the nonphysical continuum that represents the relative gravity of delinquent acts" (Sellin and Wolfgang, p. 240). The use of the power model implies magnitude estimations.

The Sellin and Wolfgang crime classification system consists of descriptions of elements of criminal events such as forcible entry, destruction of property, and assault. Certain aspects of this new classification are familiar. For example, criminal offenses are divided into two categories, category I and category II. Category I offenses are comparable to Part I offenses of the Uniform Crime Reports. That is, category I offenses are those offenses with "assumed constant reportability, violating the criminal code, known to the police, attributed to apprehended juveniles, and inflicting bodily harm on a victim and/or involving theft, damage or destruction of

property" (Sellin & Wolfgang, 1963, p.6). This is not unlike the approach taken by the Uniform Crime Reports. Part I offenses of the Uniform Crime Reports are offenses whose nature is such that they are expected to come to the attention of the police in a consistent and reliable manner. Thus, category I offenses, like the Part I offenses of the Uniform Crime Reports, are to be used to index the level of criminal activity.

One should not be misled by the apparent similarity between the Sellin and Wolfgang classification system and the Uniform Crime Reports. The Sellin and Wolfgang (1964) classification system differs from the Uniform Crime Reports in many important ways. First, the classification system focuses on the criminal event rather than on a legal label as a unit of data collection. Moreover, each criminal event is characterized by its components reflecting (1) personal injury; (2) threat and intimidation; and (3) property damaged, stolen or destroyed. A seriousness score is assigned to each component. Seriousness scores were obtained from the final stage of ratings described in a previous section. It is clear that compared to the Uniform Crime Reports, the Sellin and Wolfgang classification system allows for a finer discrimination between criminal events. For example, one is able to distinguish between, say, a robbery resulting in one victim receiving minor injuries from a robbery in which two victims needed medical treatment but were later discharged.

Second, unlike the Uniform Crime Reports, the Sellin and Wolfgang (1964) system can account for criminal events involving

multiple offenses. A robbery of some $10,000 resulting in the death of two victims, injury to three others and property damages of $2000 can be easily accommodated by this system. The index for the above event is the sum of the separate criminal acts, assuming of course that the impact of criminal acts are in fact additive. Third, since the elements of an attempted act, in general, would not be the same as the elements of a completed act, the two events would not receive the same seriousness rating. Attempted acts can therefore be differentiated from completed acts.

Sellin and Wolfgang (1964) were not the first to attempt to scale psychological variables, although they were the first to develop an empirical ratio scale of offense seriousness. Their use of seriousness ratings in a crime classification system, however, is truly innovative. Their work introduced a novel and valuable approach to the study of criminology. For this reason, a number of replications were attempted.

Normandeau (1966) performed the first replication of Sellin and Wolfgang (9164). The study took place in Canada. Two-hundred-thirty-two French Canadian students from the University of Montreal, 177 males and 55 females, were asked to rate 15 offenses. The descriptions of the offenses were similar to those used by Sellin and Wolfgang (1964) with the University of Pennsylvania students. Geometric means of the seriousness ratings were computed separately for male students, females students and for the two groups combined. All three sets of mean

seriousness ratings were found to correlate very highly with the seriousness ratings obtained from the University of Pennslyvania students. The correlation coefficients were all above 0.9. In addition, the mean seriousness ratings of male Montreal students correlated highly with the mean seriousness ratings of the Montreal females.

Normandeau regressed the University of Pennsylvania seriousness scores onto each of the three sets of seriousness scores obtained from the University of Montreal in order to examine the slope of the regression lines. Regressions which involved the University of Pennsylvania students yielded slopes which were substantially less than one, while the slope of the regression line which involved the University of Montreal males and the University of Montreal females was approximately equal to one.

Normandeau (1966) concluded that "the slopes (of the Montreal groups) in regard to Philadelphia, differ from a slope of 1 by an appreciable amount... Thus, roughly, there have been no important differences discovered within cultures, but there is some difference between cultures" (p. 176). This practice of looking at the slope of the regression lines in order to determine similarity in seriousness ratings between groups has often been attributed to Sellin and Wolfgang (1964) who stated "...we would hypothesize that in a replication of the magnitude and category scales, the scale values for offenses would be represented by (1) a slope not significantly different from

those in our study, or, minimally (2) a straight line when plotted on semilogarithmic paper" (p. 322). Sellin and Wolfgang (1964) were referring to the relationship which they obtained between their category scale and their ratio scale. Normandeau (1966), Akman, Normandeau, and Turner (1967), Reidel (1975) and others, have interpreted the statement made by Sellin and Wolfgang (1964) in a different way. Their view is similar to that of Wellford and Wiatrowski (1975), who paraphrased Sellin and Wolfgang (1964) in the following manner: "(a) At a minimum when the magnitude scores obtained in any two different study groups are plotted against each other, the display on log-log paper should be linear. The strength of this relationship can be measured using the Pearson product moment correlation. (b) At a maximum, the slopes of the lines of the two study groups plotted on log-log paper should be similar" (p. 176). Judging from the context of that paper, "similar" appeared to refer to a regression weight of one. By this definition, in order to show that a set of ratings has been replicated, one must show that (1) the log of the ratings obtained from a replication group is correlated with the log of the ratings in the original set, and (2) the slope of the regression line when the log of the ratings of one group is regressed on to the log of the ratings of the other group is equal to one. The first condition is often referred to as similarity of shape, the second condition as similarity of slope. In a later section these two conditions will be examined in more detail.

Akman, Normandeau, and Turner (1967) performed a more extensive replication in which a total of 2,745 subjects were involved. The main sample of the study consisted of 13 groups of students from 13 universities in the 10 provinces in Canada. Of the 2,384 students, 1,268 were males and 1,116 were females. Other subjects involved in the study included the following: (1) 101 English Canadian judges, (2) 52 French Canadian judges, (3) a sample of 151 officers of all ranks from the Montreal police department, and (4) a sample of 52 English Canadian white-collar workers with managerial positions in a large industrial concern.

All subjects except the judges were tested by a member of the research staff. The judges were contacted by mail. The questionnaire return rate was not reported. Each subject was given a booklet containing instructions, a set of 14 offenses to be rated, and a standard offense with a pre-assigned seriousness score of 10. Except for a small change in the description of one offense, the descriptions of the rest were the same as those used in the Normandeau (1966) study.

Following the lead of Normandeau (1966), two criteria were considered: shape and slope. Accordingly, correlation coefficients and regression weights were computed for (1) male students and female students in each province; (2) each group and the national average. The national average was obtained as a weighted average of the 10 provincial groups. The weights were chosen to reflect the relative sizes of the population in the 10 provinces; (3) each group, including the national average, and

the Philadelphia data from the study carried out by Sellin and Wolfgang (1964). All correlations coefficients were found to be 0.90 or better, and all regression weights were close to one. On the basis of these results, Akman, Normandeau, and Turner (1967) stated that "We can...firmly conclude that the scaling procedures used by Sellin and Wolfgang produce a reliable and stable index which we believe to be the best standardized measure of crime and delinquency available at present" (p. 148).

The classification system proposed by Akman, Normandeau, and Turner (1967) was almost identical to that of Sellin and Wolfgang, except for the fact that their Canadian ratings were used in place of the American ratings.

Velez-Diaz and Megargee (1971) conducted a partial replication of Sellin and Wolfgang (1964). Their sample consisted of 83 inmates of the Institute For Youthful Offenders in San Juan, and a group of 92 boys from economically deprived areas. The mean age of the offenders was 20.15 years, and the mean grade level was 6.93. The 92 members of the nonoffender group were selected from a vocational school located in the same area as the Institute. The mean age in this group was 18.2, and the average grade level was 7.4. The subjects in the Velez-Diaz and Megargee (1971) study differed from the previous studies with regard to culture, language, educational level, and socioeconomic status. In addition, the subjects in one group were youthful offenders. No offenders served as subjects in previous studies. All subjects were presented with descriptions

of 141 offenses. The offense descriptions were Spanish translations of descriptions used by Sellin and Wolfgang (1964). An 11-point rating scale was used. Means and standard deviations for each of the 141 offenses were computed. The difference in mean size of the ratings between offenders and non-offenders were tested for statistical significance with the use of t-tests. Thus, 141 t-tests were performed. Of these 141 tests, 10 were significant at the 0.05 level, and two were significant at the .01 level. Velez-Diaz and Megargee (1971), however, attributed the significant differences to chance.

Velez-Diaz and Megargee (1971) also computed Pearson product-moment correlations between ratings of the offender and non-offender groups. Using a subset of 21 offenses, the correlation was found to be 0.98. When all 141 offenses were used, the correlation dropped to 0.84. However, both coefficients were statistically significant.

Finally, the ratings of both the offender and the non-offender groups were compared to the category ratings of the Pennsylvania samples from Sellin and Wolfgang (1964). Correlation coefficients were computed and were found to be in the low 70's. An overall index of agreement, Kendall's coefficient of concordance, was also computed and was found to be 0.80. On the basis of these results, Velez-Diaz and Megargee (1971) concluded that "differences in criminality, language, culture, social class, and educational level did not result in any substantial differences in the mean ratings assigned...the

data suggest that judged seriousness of offense is a sufficiently stable unit to permit regional and cultural comparisons to be made" (p. 553).

Yet another replication was performed by Piglio (1975). The basic difference from the Sellin and Wolfgang (1964) study was that, like Velez-Diaz and Megargee (1971), Piglio (1975) also used offenders in addition to students as subjects. The Piglio sample consisted of 193 inmates from Rahway Prison, all 524 residents from Annandale Farms, a juvenile detention center, and 216 students enrolled in an undergraduate sociology class at the University of Pennsylvania. Twenty offense descriptions were chosen from the list of offenses used by Sellin and Wolfgang (1964), but were modified slightly in order to keep the language as simple as possible. As in the original Sellin and Wolfgang (1964) study, both category and ratio scales were used. Each subject participated in both scaling tasks.

Two kinds of analyses were performed on the category data. First, in order to determine the amount of agreement between the three groups with respect to seriousness of offenses, correlation coefficients were computed. All three correlations were quite large. They ranged from 0.85 to 0.95.

Second, 20 one-way ANOVAs were performed, one for each offense. The results showed that of the 20 F values computed, only two were not significant, two were significant at the 0.05 level, while the rest were all significant at the 0.01 level. Thus, in contrast to Velez-Diaz and Megargee (1971), Piglio

reported that the ratings of his groups differed from one another.

Similar analyses were performed on the data obtained from magnitude estimations. Figlio (1975) again reported that the ratings produced by the groups differed from one another. His conclusion was based on the fact that all the F-values were significant, except for one. When the magnitude ratings of the students from Figlio (1975) were compared to the data from the Pennsylvania students of the original study conducted by Sellin and Wolfgang (1964), it was found that the two sets of ratings were highly correlated ($r = 0.98$). However, since the ratings of the students from Figlio (1975) were generally about half as large as the corresponding ratings of the Sellin and Wolfgang (1964), Figlio concluded that "overall, the Penn students considered offenses as only about one-half as serious as did their counterparts ten years ago....The students have maintained the same relative regard for offenses but they have been desensitized in absolute terms." (p. 199)

For the most part, attempts to replicate the Sellin and Wolfgang (1964) study have yielded positive results. Researchers in various parts of North America have been able to produce more or less similar patterns of offense seriousness. Indeed, the results of the Puerto-Rican study and the study involving French Canadians have helped to extend the generality of the Sellin and Wolfgang technique. However, it is to be expected that any new approach would be singled out for careful scrutiny. In this

respect, the method advocated by Sellin and Wolfgang (1964) for the scaling of crime seriousness and for the classification of crime was no exception to the rule.

The power of the Sellin and Wolfgang classification system rests on a number of critical assumptions regarding the nature of their scale values. Unfortunately, many researchers are not convinced of the validity of some of these assumptions. For purposes of discussion, criticisms of the Sellin and Wolfgang (1964) approach to crime classification can be classified into four broad but overlapping categories: (1) the question of consensus, (2) issues relating to the validity of comparing seriousness ratings across different groups, (3) problem of the arbitrary modulus, and (4) issues relating to the additivity of seriousness scores when multiple criminal acts are involved.

## The Question of Consensus

To be of practical use a measure of offense seriousness must reflect the views of society as a whole. That is, if a crime classification system based on seriousness measures is to be used by law enforcement agencies and the courts for the purpose of establishing an offender's debt to society, then the establishment of the degree of seriousness of a crime is a matter which requires the consensus of society as a whole.

The issue of consensus is one which was not dealt with in a satisfactory manner by Sellin and Wolfgang (1964). Rose (1966)

criticized Sellin and Wolfgang (1964) for concluding that there was "a pervasive social agreement" about what is serious and what is not. It may be recalled that Sellin and Wolfgang (1964) used some 800 middle class subjects from the state of Pennsylvania. Many did not believe that the views of such a select group would be representative of the views of the society at large. Rose (1966) cited data obtained from the British Broadcasting Corporation which purported to show substantial differences in attitudes concerning the seriousness of certain crimes between middle and working class subjects. The research was undertaken in connection with a series of half-hour television programs called "Crime". Following the presentations, subjects were asked to name the worst crime from a list of 15 crimes. The results showed that the middle and working classes did not completely agree on the relative seriousness of the 15 crimes. While one can question the quality of the data, the issue raised by Rose (1966) is a serious one indeed. One cannot simply impose a classification system developed from a small sub-population on society as a whole without making some attempt to demonstrate that the view of society is reflected by that sub-population.

The question of consensus was also the subject of a detailed study conducted by Rossi, Waite, Bose, and Berk (1974). According to these researchers, consensus is said to exist if: (1) the correlation between the mean ratings of two subgroups in question is 0.70 or better, or (2) the regression of one group

on the other results in a zero intercept. Using a stratified sampling technique, these investigators selected 125 white and 75 black subjects from the adult population of the City of Baltimore. Descriptions of 140 offenses were developed from the Uniform Crime Reports listings by transforming broad crime categories into specific acts. Two lists of 80 offenses each were constructed. Each list had 60 unique offense descriptions, while the remaining 20 were common to both lists. During the interview, subjects were asked to rate the seriousness of the offenses on a 9-point scale. Each subject rated 80 offenses.

Correlation coefficients computed between the major subgroups showed that there were strong linear trends. For example, the correlation between blacks and whites was 0.89, between males and females was 0.94, and between high school graduates and non-graduates was 0.89. Mean seriousness ratings were also computed for subgroups classified by sex, race, and education, resulting in a total of eight subgroups. The 28 correlation coefficients computed among the subgroups ranged from 0.61 to 0.93. Black males with less than high school education were found to disagree most with the other subgroups. However, for the most part, the subjects in the eight subgroups did not vary a great deal with respect to their judgments of the relative seriousness of the crimes. Only six of the 28 correlations were in the 0.60s, the rest of the correlations were well above the minimum required for consensus as defined by Rossi, Waite, Bose, and Berk (1974). Rossi, Waite, Bose, and

Berk (1974) also reported an intercept of -.39 when ratings obtained from white subjects were regressed on ratings of black subjects, and an intercept of -.80 when male ratings were regressed on female ratings. This, according to these researchers, indicated that blacks tended to see crimes as more serious than whites and women tended to see crimes as more serious than males. Nevertheless, they concluded that overall, there was an impressive amount of consensus among the subgroups. It should be noted that some subgroups means were based on sample sizes as small as seven. Caution should therefore be exercised in the interpretation of their results.

Walker (1978) presented 11 crime descriptions to some 690 subjects to be rated. The subjects were classified into groups according to social class and sex. Three scaling techniques were used: the method of paired comparisons, an 11-category interval scale, and a ratio estimation procedure. Each subject participated in all three scaling tasks sequentially in the order listed above. Analyses of variance were performed for each one of the offense descriptions, first using category data, and subsequently using magnitude estimations. Since a number of the F-tests were significant, Walker (1978) concluded that "there are some significant differences between social classes and sexes as to how they assess the relative seriousness of certain offenses" (p. 363).

Hsu (1973) studied crime seriousness in Taiwan. Her sample consisted of 239 male and 60 female students from the National

Taiwan University, a group of 248 policemen, and a group of judges (sample size not mentioned). Fourteen index offenses from Sellin and Wolfgang (1964) were used. The subjects were asked to rate the offenses with respect to a standard offense with an assigned scale value of 10. The correlations between the ratings of the three groups of subjects were all better than 0.90. The correlation between male students and female students was also computed and was found to be 0.84.

When the Hsu ratings were compared to the Philadelphia ratings of the original Sellin and Wolfgang (1964) study, and with the ratings of Canadian students from Akman and Normandeau (1968), she found that the Taiwan ratings correlated 0.95 with the Philadelphia ratings and correlated 0.90 with the Canadian ratings. The slope of the regression of Taiwan ratings on Canadian ratings was found to be 0.53, and the slope of the Taiwan ratings on the Philadelphia ratings was 0.60. Hsu (1973) concluded that her results met the similarity of shape criterion, but not the similarity of slope criterion.

Judging by these results, the issue of consensus is far from settled. For example, Normandeau (1966), and Walker (1978) found evidence to indicate that there were group differences. Akman, Normandeau, and Turner (1967) on the contrary, found no such evidence. Hsu (1973) obtained conflicting results, and so did Rossi, Waite, Bose, and Berk (1974). Closer inspection of these studies leads to the conclusion that at least some of the confusion may be attributed to the fact that a number of

different criteria were used to determine "similarity". Walker (1978), for example, examined mean group differences. Rossi, Waite, Bose, and Berk (1974) defined consensus based on the size of the correlation coefficient, and on a zero intercept. Figlio (1975), and Normandeau (1966), used the correlation coefficient. Normandeau (1966), Akman and Normandeau (1967); and Akman, Normandeau, and Turner (1967) used a unit slope as one measure of agreement or consensus. Since these criteria are not completely equivalent, it is not surprising that researchers have not been able to agree with one another regarding the generalizability of the Sellin and Wolfgang (1964) ratings.

## Comparison of Seriousness Scores Between Groups

Researchers attempting to compare seriousness ratings obtained from different groups, especially with ratings obtained from different studies are faced with numerous challenges. The lack of consensus on the choice of a criterion is perhaps the most serious challenge. Satisfactory resolution of this problem must precede the search for solutions to issues such as the question of consensus. Given the fact that researchers have not used the same criterion in their attempts to compare seriousness ratings between groups, it is appropriate, then, to examine the strengths and weaknesses of each of these criteria in some detail.

Correlation and difference between overall groups means.

The Pearson-product moment correlation coefficient is a commonly used measure of linear relationship or agreement. Given two sets of ratings, the correlation coefficient is sensitive to differences in the relative distances of the offenses on the seriousness continuum between two sets of ratings. The correlation coefficient is not sensitive to overall mean differences in the ratings between the sets, nor is it sensitive to differences in units of measurement. In other words, so long as the relative distances between ratings are more or less proportional in the two sets of ratings, the ratings would be correlated. The overall magnitude of the ratings would not affect the size of the correlation. Since differences in magnitude are also relevant in the consideration of whether or not two sets of ratings are similar, the existence of a strong linear trend alone is not sufficient to demonstrate that the two sets of ratings in question are equivalent. By the same token, the lack of an overall mean difference between groups is not sufficient to indicate consensus, since the relationship between the two groups may be non-linear.

When mean difference is used as a measure of consensus, there is an additional consideration. The experimenter must ensure the the same metric is used. The comparison of mean ratings based on different metrics obviously does not make much sense. This is illustrated in a study conducted by Figlio (1975) discussed in a previous section. It may be recalled that the

ratings obtained by Figlio (1975) correlated highly with those of Sellin and Wolfgang (1964). Inspection of the magnitude of the ratings, however, led Figlio (1975) to conclude that his students considered the offenses to be only half as serious as Sellin and Wolfgang's Pennsylvania students. However, without adequate information regarding scaling procedures, it is difficult to tell whether the differences are real or are merely due to a difference in unit of measurement. Unless specific information is provided regarding the unit of measurement the slope of the regression is likely to be very difficult to evaluate.

Slope of the regression line. The slope of the regression line has also been frequently used as a criterion for measuring similarity in seriousness ratings. Akman and Normandeau (1967), for example, asserted that if two groups agree as to the relative seriousness of offenses, then the slope would be equal to one. In other words, the slope of the regression line contains the necessary information to determine the degree of similarity between sets of ratings. The real situation however, is somewhat more complex. Unlike the correlation coefficient, the slope of the regression line is sensitive to differences in unit of measurement. The slope is a function of the degree of linear relationship between two sets of ratings and the ratio of their respective standard deviations. More specifically, slope is equal to r(sd1/sd2), where sd1, and sd2 are the standard deviations of the ratings in set 1 and set 2 respectively, and r

is the correlation. Since the standard deviation is proportional to the unit of measurement, a change in the unit of measurement changes the slope of the regression line. For example, suppose the length of a number of sticks were measured (in inches) on two different occasions. If the measurements are assumed to be accurate, the two sets of measurements would have a correlation of one, and a slope of one. However, if the first set of measurements were in centimeters and the second set were in inches, the two sets of measurements would still be completely equivalent to one another. The correlation coefficient would be equal to 1, but the regression of the metric measures on the English measures, would not result in a slope equal to one. Instead, the regression would yield a slope equal to 2.54, approximately, which incidentally is the ratio of metric versus English measures. Clearly, the unit slope criterion is not applicable in all cases. One solution is to ensure that the same unit of measurement is used. In ratio estimation, for example, the use of a common "standard" offense description with a pre-assigned value in fact serve to define the unit of measurement. Since in ratio measures the origin is fixed, defining a second point on the scale defines the unit of measurement. Unfortunately, researchers have not always insisted on using a comparable unit of measurement. As a result, it is often difficult to relate the slope of the regression line to degree of consensus.

Non-zero intercept. Rossi, Waite, Bose, and Berk (1974) used a non-zero intercept as a criterion for overall mean difference in ratings between groups. For example, when the ratings of white subjects were regressed on the ratings from black subjects, Rossi, Waite, Bose, and Berk (1974) obtained an intercept of -.39. The intercept of the regression equation of male ratings on female ratings was -.80. Rossi, Waite, Bose, and Berk (1974) therefore concluded that blacks tended to rate each crime as more serious than whites by 0.4 units and females tended to rate crimes more seriously than males by 0.8 units. The problem with this approach is that a non-zero intercept does not always indicate an overall mean difference. The relationship is somewhat more complex than that envisaged by Rossi, Waite, Bose, and Berk (1974). The intercept is a function of the overall mean difference, the degree of linear relationship between the two sets of ratings and the ratio of their respective standard deviations. A non-zero intercept therefore tends to be difficult to interpret.

In their study, Rossi, Waite, Bose, and Berk (1974) used a 9- point interval scale. If crime seriousness is measured on a ratio scale, as is the case in many of the studies in this area, a non- zero intercept takes on quite a different meaning. Since the zero point in ratio scales is not arbitrary, the regression line must pass through the joint origin (0,0) if the linear relations model is to hold. A non-zero intercept in this case likely indicates that something is amiss with the fit of the

ratio scale model to crime seriousness data.

Arbitrary modulus. Comparison of seriousness ratings between groups is certainly not a simple matter, especially when crime seriousness is measured on a ratio scale. When crime seriousness is measured on a ratio scale, other complications must be dealt with. In magnitude estimation situations in psychophysics, for example, in judgments of brightness of a light source or loudness of a tone, the size of the numbers used is frequently determined by the subjects. It is fairly obvious that in such a situation, comparisons of absolute magnitude between subjects is quite useless. This is because ratings produced by direct ratio estimation are not necessarily tied to any specific unit of measurement, or modulus. The term modulus is synonymous with unit of measurement, and is used interchangeably in the present context.

In a ratio estimation situation where a subject is shown two sticks and is asked to first assign a number to stick A, and then to assign a second number to stick B such that the values reflect the number of times stick B is longer than stick A, the subject is not restricted in his choice of numbers. If stick B is in fact twice as long as stick A, the subject is free to assign, say, 5 to stick A and 10 to stick B. Alternately, he can assign 1 to stick A and 2 to stick B. In addition, whatever numbers the subject eventually chooses, one can always apply a transformation of the form $y=ax$ to the ratings, where $y$ is the new scale value, $x$ is the original scale value and $a$ is any real

number constant. Clearly, the absolute values of the numbers are not of any interest in and of themselves, even though the numbers are meaningful because they do correctly reflect the relative amounts of the attribute under study. This indeterminacy has proved to be troublesome for the many researchers who have attempted to compare magnitude estimations across various groups of subjects (Akman & Normandeau, 1967; Figlio, 1975). For example, results of a study undertaken by Kvalseth (1980) to compare ratings obtained from Norwegian students with those obtained by Akman and Normandeau (1967) is difficult to assess. Kvalseth (1980) did not use a standard for reference, nor did he fix the unit of measurement. In situations where subjects are allowed to use any number they wish, the absolute size of the ratings is not of any interest. In this respect, the use of such statistical procedures as the t-test and analysis of variance in these situations is not appropriate.

Pease, Ireson, and Thorpe (1975) recognized this indeterminacy problem. Their solution was to transform magnitude scores by subtracting from each rating the median of the respective groups, and dividing through by the average absolute deviation from the median. In effect, they substituted one problem for another. Their approach amounted to the assertion, without proof, that the median seriousness for all groups are equal.

In spite of the fact that most magnitude estimation procedures yield scale values which are only determined to

within some linear transformation of the form $y = ax$, or in the logarithm form $\log(y) = \log(a) + \log(x)$, researchers continue to routinely compare seriousness estimations across groups. A procedure which can overcome this problem of the arbitrary modulus should therefore be a welcome addition to the tools available to researchers working in this area.

## The Question of Additivity

A central part of the Sellin and Wolfgang (1964) crime classification system is the additivity of seriousness scores. Their system focuses on the criminal event as the basic unit of analysis rather than on categories defined according to legal labels. Since a criminal event may involve one or more elements, where the elements are either distinct criminal acts or factors which aggravate it, each criminal event can be scored as the simple sum of the sericusness of its individual elements. Whether crime seriousness scores can be added in this manner is open to debate. Sellin and Wolfgang (1964) did not conduct a specific test for non-additivity. Many investigators believe that the Sellin and Wolfgang system adequately reflects the seriousness of criminal events regardless of whether the event involved a single offender committing a single offense or a single offender committing multiple offenses or multiple offenders committing single or multiple offenses (Wellford & Wiatrowski, 1975). Others, however, believe that the seriousness

of complex criminal events cannot be obtained as a simple sum of the seriousness of the individual elements (Pease, Ireson, & Thorpe, 1974; Rose, 1966). These latter researchers believe that the perceived seriousness of a criminal event depends on the particular mix of individual criminal events. In other words, they believe that depending on the particular mix of events, the whole is more or perhaps even less than the sum of the individual parts.

On the subject of additivity, Rose (1966) commented: "the important assumption that one can merely add scores for complex events has not been tested by the authors....Although the scoring system was designed so that it could in particular deal with complex events no such complex event was included in the list of 141 offenses" (p. 421). In an attempt to shed some light on this issue of additivity, a number of studies were undertaken.

Pease, Ireson, and Thorpe (1974) conducted an experiment to test the assumption of additivity. Their sample consisted of 147 students enrolled in evening classes in the Manchester area. Of the 147 students, 92 were females, and 55 were males. Two forms were devised each containing four pairs of offenses, of which two pairs were fillers. The students were presented with either form A or form B. The task was to determine if the second offense was (1) less serious, (2) equally serious, (3) one and a half time as serious, (4) twice as serious, (5) two and a half times as serious, (6) three times as serious, or (7) more than

three times as serious, compared to the first.

The results showed that over all four pairs of offenses used in the study, roughly 32 per cent of the ratings were in the twice as serious category. Some 48 per cent of the judgments were classified in the less than twice category, and 20 per cent were judged to be more than twice as serious. This, according to Pease, Ireson, and Thorpe (1974), contradicted the assumption of additivity, since only some 32 per cent of the responses were in agreement with the prediction of Sellin and Wolfgang (1964).

Pease, Ireson, and Thorpe (1974) were criticized by Wellford and Wiatrowski (1975) who pointed out that their results cannot be used to further our understanding of the additivity issue, because of a "fatal methodological error". In three of the four pairs of offenses, the two elements of the event were separated in time. In one case the two elements were separated by "a few days", in another by "later", and in the third by "soon after". These critics suggested that if the ambiguity caused by the time dimension was taken into account the data collected by Pease, Ireson, and Thorpe would in fact support the additivity assumption.

Wagner and Pease (1978) replicated the study by Pease, Ireson and Thorpe (1974). The same procedure was used, however, the four sets of offense descriptions were changed so that there was no temporal separation between the multiple criminal acts. The results still showed that overall, only about 18 per cent of the judgments were classified as "twice as serious".

Wellford and Wiatrowski (1975) collected data from 118 students from the Florida State University (F.S.U.). Two sets of offense descriptions were compiled. The first set consisted of 37 descriptions of simple offenses, all taken directly from The Measurement of Delinquency, a book authored by Sellin and Wolfgang and published in 1964. The second set contained 21 descriptions of complex offenses. The offenses were similar to the ones in the first set, but were modified so that multiple criminal acts were involved. For example, offense description number three in set one read: "The offender robs a person of $5 at gunpoint. No physical harm occurs" (p. 185). In the second set, the offense read: "The offender robs two persons of $5 at gunpoint. No physical harm occurs" (p. 187). The changes varied from one offender committing an offense against two persons to three offenders each committing two offenses. A magnitude estimation procedure was used, and each student rated the offenses in both sets.

Since the complex criminal events used by Wellford and Wiatrowski (1975) were made up of a number of discrete criminal acts, the seriousness of these events could be derived indirectly, by adding up the seriousness scores of the individual criminal acts. Two such indirect scales were derived. One used scale values from students at the F.S.U. Since the crime descriptions were taken from Sellin and Wolfgang (1964), the second indirectly derived scale values were obtained by adding up the scores from the Sellin and Wolfgang (1964) study.

The correlation between the directly scaled complex crimes and the P.S.U. indirectly scaled complex crimes was 0.97. The correlation between the directly scaled complex crimes and the Sellin and Wolfgang (1964) indirectly scaled crimes was 0.90. The substantial size of the correlations led Wellford and Wiatrowski (1975) to conclude that: "the additivity assumption is strongly supported. While the absolute values of seriousness are surely changing, the results of our research lead us to the conclusion that the minimum condition of additivity replication is supported" (p. 183) Visual inspection of the data provided by Wellford and Wiatrowski (1975) showed that direct and indirect scale values for the complex crimes of the P.S.U. students were generally of the same magnitude, but the indirect scale values derived from Sellin and Wolfgang (1964) were only about half as large as those obtained from the P.S.U. students.

A conclusion of strong support for additivity is perhaps not totally justified, since the question of additivity cannot be resolved by the use of the correlational technique alone. The magnitude of the ratings must also be taken into consideration. If crime seriousness is additive, then directly scaled and indirectly scaled multiple criminal events should yield seriousness scores of approximately the same magnitude. The ratings for some crimes differed by a fairly large amount, even within the P.S.U. sample, indicating that perhaps others factors need to be taken into account. Without additional studies, the question of additivity cannot be properly answered.

The large difference in ratings between the scores derived from Sellin and Wolfgang (1964) and the P.S.U. scores cannot be easily explained away. It is perhaps a further illustration of the problems which stand in the way of meaningful comparisons of magnitude scores obtained from different studies.

## III. Purpose and Scope of the Present Study

There is little doubt that the crime seriousness scale and crime classification system pioneered by Sellin and Wolfgang (1964) appeal to many criminologists. This is evidenced by the fact that the seriousness index developed by these authors continues to be popular among researchers (Sheley, 1980). However, the lack of agreement concerning issues such as consensus, and additivity, and on how to deal with the problem of the arbitrary modulus points to the need to re- examine some of these issues.

The purpose of this study, therefore, is to extend our current understanding of crime seriousness scaling. With this in mind, the following three issues will be examined:

(1) The goodness of fit of the power function model to the subjective crime seriousness data.

(2) The arbitrary modulus, or arbitrary unit of measurement problem and how the problem might be overcomed.

(3) The issue of additivity of seriousness scores.

## Goodness of Fit

When magnitude estimation procedures are used in psychophysics, the adequacy of the power function model is frequently assessed by examining the correlation coefficient between the logarithm of the physical stimulus and the logarithm of the response. In the scaling of psychological variables where the values of the stimuli are generally not known, researchers have frequently resorted to the examination of the correlation coefficient between responses obtained by Fechner's or Thurstone's category scaling procedures and the logarithm of magnitude estimations. Stevens (1966) suggested the use of this procedure. He stated that "if the same relations that have been shown to obtain in sensory psychophysics among the three general kinds of measures can also be shown to characterize the comparable scales created with nonmetric (psychological) stimuli, added confidence attaches to the outcome" (p. 532).

The use of the correlation coefficient for model testing has been criticized by several investigators including Birnbaum (1973, 1974) and Anderson and Shanteau (1977). These investigators have pointed out that any monotone function will have a large linear component. Good (1972) demonstrated that for some nonlinear monotone functions the correlation coefficient can approach one. Since all monotone increasing functions tend to have high correlations, a high and significant correlation would not necessarily indicate that the power model or some

other model is appropriate for the data.

The issue as to whether or not the so called ratio scales of crime seriousness in fact possess ratio properties has never been explicitly studied. For example, no one has attempted to assess the goodness of fit of the power model to the subjective crime seriousness data. Nevertheless, this issue is a fundamental one for any crime classification system which uses scale values obtained by the use of the power model. For this reason, a goodness of fit test is proposed in this study, and the fit of the power model to crime seriousness ratings will be tested.

Arbitrary modulus

Sellin and Wolfgang (1964) chose to use ratio scales instead of category scales because as they put it "magnitude estimation scale values are a product of the rater rather than the experimenter, and as such have an inherent validity that cannot be claimed for the imposition of a fixed range of category values by the experimenter on the rater's judgment....the freedom in the range of possible responses available by the magnitude estimation technique provides intrinsically more information about the raters' judgments than the severely limited categories" (p. 272). Unfortunately, the price for this flexibility is that the unit of measurement used is unique to each rater. Among other things, the presence of the

arbitrary modulus problem implies that strictly speaking, averaging ratings over raters is not appropriate. Since it appears that most researchers prefer the use of ratio scales, a solution must be found. In this study, a method to overcome the arbitrary modulus problem is proposed.

## Additivity

In general, the analyses used in studies concerned with additivity of crime seriousness ratings lack sophistication. More often than not, conclusions were based on visual inspection of the data and on the inspection of such things as correlation coefficients and slope of regression lines. In this respect, new approaches need be explored, and certainly, more rigorous data analytic techniques need to be employed. An example of a study which used a sound technique is one conducted by Gottfredson, Young, and Laufer (1980), who set out to investigate the additivity assumption from a new perspective. They developed descriptions of five offenses, involving theft, check fraud, burglary, robbery, and vandalism. Each of the five offenses appeared ten times, each time with a different amount of monetary or property loss. The ten different amounts were: $5, $10, $20, $50, $100, $500, $1,000, $5,000, $10,000. The subjects were 159 undergraduate and graduate students enrolled at John Hopkins University. All subjects rated all 50 crime descriptions using an 11-point category scale. A two-factor repeated-measures

analysis of variance design was used. P-values for the two independent variables, crime type and monetary loss, were significant ($p < 0.001$). Of greater interest is the fact that the interaction between crime type and monetary loss was also significant ($p < 0.001$). Inspection of the graphs of the power functions relating monetary loss and judged seriousness for the five offenses (Gottfredson, Young & Laufer 1980), showed that the offense of robbery appeared to be primarily responsible for the observed interaction effect. It is to be noted that of the five offenses, robbery was the only offense which involved any degree of confrontation between offender and victim, and thus, was of a more serious nature compared to the other crimes.

In a second experiment, Gottfredson, Young, and Laufer (1980) added two other types of crime to the original list of five. These were rape and robbery involving the death of a victim. The subjects were 158 students from Rutgers University. An analysis of variance again showed a significant interaction term. Furthermore, inspection of the graphs showed that the slope of the line relating seriousness and monetary loss appeared to be less steep for the more serious crime of robbery and rape, than crimes such as theft or check fraud. Thus, given the offense of rape, monetary loss seem to add very little to the seriousness of the criminal event.

The implications of this result are quite clear. The perceived seriousness of a complex crime is not only a function of the seriousness of the individual criminal act, but is also a

function of the interaction between the criminal acts which form part of the total criminal event. One can for example, speculate that in a robbery in which two of the victims are killed, the robbery itself might appear to be less serious than usual, when seen in the light of the two homicides. If this is the case, an additive model might not describe the phenomenon accurately. In general, such a model might tend to overestimate the seriousness of complex offenses.

The use of analysis of variance to test for additivity is vastly superior to other previously used techniques. Regrettably, analysis of variance techniques are not appropriate with the usual ratio scaling technique since scale values are determined only to within some linear transformation of the form $y=ax$. Gottfredson, Young, and Laufer (1980) used an 11-point category rating procedure. However, since the ratio scaling technique proposed in this study overcomes the arbitrary modulus problem, the additivity of ratio scales of crime seriousness will be assessed with the use of the analysis of variance.

## Theoretical Considerations

In their development of a crime seriousness scale, Sellin and Wolfgang (1964) made use of the power function. Stevens' power law stipulates that subjective magnitude is a power function of the stimulus magnitude (Stevens, 1957). Thus if r is the subjective magnitude and x represents some attribute of the

object being scaled, the function relating the two is of the form:

$$r = ax^{b} \quad (1)$$

The subjective magnitude r is determined by the value of the stimulus x raised to some power b. The exponent b is said to be characteristic of the phenomenon under study. All this is determined to within some arbitrary multiplicative constant a. Equation 1 can be expressed in another form. Taking logarithms on both sides of the equation, we get:

$$\log(r) = \log(a) + b \log(x) \quad (2)$$

Equation 2 is the familiar equation to a straight line, where b is the slope and log(a) is the intercept. In a magnitude estimation situation with no prescribed modulus, the intercept is a function of the subject and his particular preference for the use of certain number ranges. Thus, log(a) can be viewed as some 'mean' level which may differ from individual to individual. The standard procedure in crime seriousness research uses magnitude estimations directly as the data for analysis. It is obvious that comparisons of these magnitudes by computing the geometric means of y or the arithmetic means of log(y) across individuals or across groups (if aggregation of individual data could be justified) yield results which may be less than meaningful.

## Procedure for scaling

The method developed here involves the presentation of stimuli for comparative judgment in a paired comparison situation. Comrey (1950) proposed a similar method. In the Comrey experiment subjects had the task of dividing 100 points between two stimuli so as to reflect the ratio of their respective attributes. Thus, stimuli were presented two at a time. All possible pairs of stimuli were presented. In the present study, each subject will be presented with a single pair of stimuli, chosen from a list of eight stimuli (Appendix A). One of the stimuli, break and enter, will be used as a baseline condition. The other seven stimuli all include break and enter in addition to other crime descriptions. The task of the subject is to assign numbers to the pair of stimuli presented to him so as to reflect the ratio of the magnitude of the attribute under study. The method proposed here differs from that of Comrey in that a slightly different kind of restriction is placed on the size of the numbers the subjects can use. In this study, the subject is asked to assign a 10 to the least serious of the two crimes. This number is then used as a "standard" in the paired comparison task. All possible pairs of stimuli are presented. With a total of eight stimuli, the number of combinations taken two at a time is 28. Thus, there were 28 pairs of stimuli.

The presentation of pairs of stimuli produces data which can be used to overcome the indeterminacy in the estimated magnitudes, inherent in most ratio scaling procedures. In this case, the ratio of the estimated magnitudes is used instead of the magnitudes themselves as in previous studies. If r(i,j) and r(i,j') are estimated seriousness scores obtained by comparative judgment from subject (i) on stimuli (j) and (j'), where j > j'.

$$(3) \qquad r(i,j) = a(i) * x(j)^{b}$$

$$(4) \qquad r(i,j') = a(i) * x(j')^{b}$$

then,

$$(5) \quad (r(i,j)/r(i,j')) = (a(i) * x(j))^{b}/(a(i) * x(j'))^{b}$$

$$= (x(j)/x(j'))^{b}$$

The multiplicative constant conveniently disappears. The logarithm form is:

$$(6) \quad \log(r(i,j)/r(i,j')) = (\log(a(i)) + b \log(x(j))) - (\log(a(i)) + b \log(x(j')))$$

$$= b\ (\log(x(j)) - \log(x(j')))$$

Again, the intercept ($\log(a(i))$) disappears. Equations 5 and (6) draw attention to the fact that the data in the form of $\log(r(i,j)/r(i,j'))$ or $\log(r(i,j))-\log(r(i,j'))$ are adequately determined, and thus, allow the use of analyses such as the t-test and analysis of variance to test for hypotheses concerning

the degree of seriousness of crimes under the various experimental conditions.

## Testing for Interaction Effects

The existence of non-additivity in offence seriousness scores can be conveniently tested by performing an analysis of variance. Since each crime description is paired with every other description, and since there are two levels to each description, present or absent, the data can be represented by a 2 X 2 X 2 factorial design (Appendix B). A statistically significant interaction term would, in this case, indicate that offence seriousness scores are not additive.

The ordinary linear model is represented by the following equation:

$$y = Xb + e \tag{7}$$

The model for the present study can be represented by equation 8,

$$\bar{r}(j,j') = CXb + e \tag{8}$$

where $\bar{r}(j,j')$ is a vector of mean observed (log) differences. That is, $r(j,j')$ is obtained by averaging $\log(r(i,j)) - \log(r(i,j'))$ over subjects. The usual vector of regression weights is b, and e is a vector of random errors. X is an 8 by 8 design matrix representing 3 independent variables, two-way and

three-way interactions of the ordinary linear model for 3 independent variables each having two levels: zero, and one.

(9)

| | Effect | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | B | V | T | A | VT | VA | TA | VTA | | |
| | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | B | |
| | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | V | |
| | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | T | |
| X = | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | A | Crime |
| | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | VT | Type |
| | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | VA | |
| | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | TA | |
| | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | VTA | |

The rows of the X matrix represent crime descriptions and the columns represent the effects. The absence of an effect is denoted by zero, and its presence is denoted by one. Thus, for example, the baseline condition, break and enter, vandalism, assault, and a two-way interaction effect, vandalism and assault all contribute to the overall effect of crime description 6 (row 6).

Since $\bar{r}(j,j')$ is actually a vector of differences obtained by comparative judgments, that is $\bar{r}(j,j')$ is obtained by averaging $\log(r(i,j)) - \log(r(i,j'))$ over subjects, the contrast matrix C is therefore used to modify the basic design matrix X in order to reflect the subtractions. The form of the C matrix is shown in equation (10).

$$
(10) \quad C = \begin{array}{cccccccc}
B & V & T & A & VT & VA & TA & VTA \\
-1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
-1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
-1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 1
\end{array}
$$

Row one of the contrast matrix C, for example, represents the comparative judgments involving the baseline and vandalism. The C matrix is used as a means of adjusting the design matrix X to indicate for each comparison, which effect is in or out. Thus, the matrix CX consists of zeros, ones, and minus ones, and indicates whether an effect is absent, present or is subtracted from the overall effect. The regression weights represent the effect of the corresponding conditions.

Inspection of equation 8 reveals that the first three regression weights have an additional function. In the present framework, the first three regression weights are also the log of the estimated scale values for vandalism, theft, and assault, respectively. Since the estimation procedure produces relative seriousness values, the CX matrix has only seven columns instead of eight. This simply means that the log of the scale value for

the baseline condition is arbitrarily set to zero. The scale values for vandalism, theft, and assault are thus estimated relative to baseline value of zero, in log units.

## Goodness of Fit Test

The replication of each paired comparison across a number of subjects provides a means for testing the goodness of fit of the power function model to the subjective seriousness data. The goodness of fit test proposed here involves the comparison of two models representing the data.

$$r(i,j,j') = t(j,j') + e(i,j,j') \quad (11)$$

$$r(i,j,j') = [S(j) - S(j')] + e(i,j,j') \quad (12)$$

Equation 11 represents an unconstrained model where the observed score r(i,j,j') from subject i rating a pair of stimuli j and j', is equal to the true score t(j,j') plus error e(i,j,j'). Equation (12) represents a constrained model. In this model the variability associated with the residual error consists of two components. One component is a constant for all subjects in the (j,j') group, but may vary across groups. The second component represents uncontrolled variability of the ith score from the (j,j') group mean. The errors, e(i,j,j'), are assumed to be normally distributed with mean equal to zero and variance equal to some value sigma. The test of "goodness of

fit" consists of a comparison of the error variances obtained from the two models. An estimate of the error variance of the unconstrained model can be obtained from SSE(1), which is obtained by pooling the sum of squared deviations of the responses r(i, j, j') about their group means. SSE(1) and the degrees of freedom for SSE(1) are given by the following equations:

$$(13) \qquad SSE(1) = \sum_{j} \sum_{j'} \sum_{i} [r(i,j,j') - \bar{r}(j,j')]^2$$

$$(14) \qquad df(1) = \sum_{j} \sum_{j'} [N(j,j') - 1]$$

The error variance for the constrained model can be obtained from SSE(2) (equation 15). S(j) and S(j') are computed scale values for stimuli j and j'.

$$(15) \quad SSE(2) = \sum_{j} \sum_{j'} \sum_{i} [r(i,j,j') - (S(j)-S(j'))]^2$$

SSE(2) can be partitioned into two parts (equation 16). This conceptualization of SSE(2) is particularly useful, since SSE(3) can be viewed as the sum of squared deviations due to "lack of fit". SSE(3) and the degrees of freedom associated with SSE(3) are given by equations 17 and 18.

$$(16) \qquad SSE(2) = SSE(1) + SSE(3)$$

$$(17) \quad SSE(3) = \sum_{j} \sum_{j'} N(j,j')[(S(J)-S(J')) - \bar{r}(j,j')]^2$$

$$(18) \quad df(3) = [p(p-1)/2] - (p-1) = (p-1)(p-2)/2$$

If the power model is appropriate for the subjective seriousness data, both quantities SSE(3)/df(3) and SSE(1)/df(1) provide independent estimates of the error variance. The quantity [ SSE(3)/df(3) ]/[ SSE(1)/df(1) ] is distributed as F with df(3) and df(1) degrees of freedom. This quantity can therefore be used as a statistical test for the goodness of fit of the model. To the extent that the power model lacks proper fit, SSE(3)/df(3) will be inflated, and the computed F will be large.

## IV. Methodology

### Subjects

Subjects for the study were selected from the undergraduate student population at Simon Fraser University. The students were enrolled in Social Sciences courses for the 1982-1983 academic year. A total of 1082 students, 521 males and 561 females, served in the study. The median age of male subjects was 21.7, and the median age for female subjects was 20.8. Most of the subjects were single (85%). Some 3.1 per cent were living under common law, and 7.4 per cent were married. Legally separated or divorced students made up 3.3 per cent of the sample, and 1.2 per cent did not respond to the item. Just over one fifth of the sample, 22.9 per cent, indicated that they had been the victim of a crime in the past year.

### Independent variables

Four independent variables, break and enter, theft, vandalism, and assault were used in the study. The choice of these independent variables was based on several considerations. Descriptions of very serious crimes were excluded from the study. It was felt that if the power model is an appropriate

model, then it would surely be appropriate for crimes of low to moderate seriousness. The selection of low to moderately serious crimes, therefore, tends to bias the results in favor of the model. If the power model can be shown not to fit under these conditions, the results would be all the more meaningful in the broader context of crimes in general.

Another major concern of this study is the additive effects of crime on the perception of seriousness. The ability to mix various crime descriptions in order to form plausible incidents is therefore important. The selection of the four crimes, break and enter, vandalism, theft, and assault, made it relatively easy to construct mixed incidents. Descriptions of criminal events were generated using one or more of the four crime types. One independent variable, break and enter, served as the baseline condition. That is, descriptions of criminal events always included break and enter. The other three types, theft, vandalism or assault, were either present or absent. This arrangement can be conceptualized in terms of a 2 by 2 by 2 factorial design (Appendix B) requiring eight descriptions of criminal events. A complete list of all eight descriptions can be found in Appendix A. For convenience, each crime description is identified by a six-character label (Appendix A). For example, A0T1V1 identified the event as a break and enter resulting in theft (T1) and vandalism (V1) but no assault (A0). A0T0V0 indicated that the event was simply a break and enter, no theft (T0) or vandalism (V0) or assault (A0) took place. The

eight items in Appendix B formed the basis of the questionnaires which were presented to the subjects. Each of the eight descriptions was paired with every other description, resulting in a total of 28 different pairs of crime descriptions. Each questionnaire, however, contained only a single pair of crime descriptions.

## Procedure

The questionnaires were administered to subjects, on a random basis, during their regular class time. Each subject was presented with the descriptions of a pair of criminal events. All possible pairs of the events were presented. Each subject, however, responded to only one pair. The subjects were required to directly estimate the seriousness of the two crime events relative to one another by assigning numbers to each of the descriptions. This procedure is similar to that used in many previous crime seriousness scaling studies. Prior to the administration of the questionnaires the subjects were instructed to read the instructions which were attached to each questionnaire. The instructions were as follows:

> On the following page are descriptions of two crime events. Your task is to tell how serious the two events are relative to one another. Assign a value of 10 to the crime which you think is less serious. Then assign a number to the other crime so that the number reflects the seriousness of this crime relative to the less serious one. Confirm your decision by asking yourself whether the second crime is in fact that many times more serious than the one assigned a 10. For example, if you

feel that crime B is less serious than crime A, you assign 10 to crime B. Now, looking at the two crimes again, you feel that crime A should be assigned a score of 25. Is crime A two and a half times as serious as crime B? If it is, then crime A should receive a score of 25. If you don't feel this is quite right, reconsider and make another assignment. You can use whole numbers, or decimals. Just make sure that the assignments you make are proportional to the seriousness of the crimes as you perceive them.

Since each subject responded to a pair of crime descriptions, it was necessary to counterbalance for a possible order effect. In this regard, the order in which the events appeared was reversed for half of the subjects. This procedure in effect doubled the number of pairs of questionnaires from 28 to 56.

The questionnaires were organized into sets prior to distribution. Each set contained all 56 pairs of items. All questionnaires in a set were distributed before another set was placed in circulation. This procedure was adopted so as to ensure that no single group of subjects responded to an excessive number of a particular pair of items. To some degree, this approach also ensured that approximately the same number of subjects responded to each of the 56 questionnaires.

The research plan called for the collection of data for the first 20 male subjects and the first 20 female subjects for each of the 28 cells of the design. The questionnaires were therefore divided into two equal stacks. One stack was earmarked for distribution to male subjects and the other was marked for distribution to female subjects. Throughout the data collection phase strict accounting of the questionnaires was kept. Any

questionnaires that were spoiled were replaced immediately. The number of spoiled questionnaires were recorded.

The additional work which resulted from the adoption of this rather tedious procedure was well warranted since it resulted in reasonably well balanced cell sizes, and yielded approximately the same number of male and female subjects. This should, in general, make data analysis somewhat less complex.

## V. Results and Discussion

### Preliminary Analyses

Tests for order difference. The data were first examined to determine whether or not the two forms used in the study were comparable. In order to ensure that the results would be valid whether a sex difference was present or not, tests for an order effect were performed independently for male and female subjects.

The null hypothesis of no overall difference between the means of the two forms (orders), across all 28 cells of the design, can be tested with the use of an F-test (Equation 19).

(19) F = (mean square order/mean square within)

The mean square for order is given by:

$$(20)\sum_{j}^{28}((\bar{r}(1,j) - \bar{r}(2,j))^{2} / ((1/n(1,j))+(1/n(2,j)))) / (df)$$

where n(1,j) and n(2,j) are the sample sizes for form 1 and form 2 respectively.

The mean square within is the error term from the two-way, cell by form analysis of variance. The presence of an order effect would tend to inflate the 'mean square order' and thus result in a significant F.

Fortunately, for the study, the F-tests did not reach statistical significance. For male subjects, $F(27,465) = 1.073$, $P > .05$, and for females, $F(27,505) = 0.795$, $p > .05$. Normally, the results of these F-tests would have provided reasonably strong evidence that there was no order effect in either of the two sex groups. However, preliminary inspections of the data revealed the possible existence of heterogeneity of cell variance. The homogeneity of variance assumption underlying the F-test may therefore have been violated. The exact probabilities of the F values would be all but impossible to assess. Nevertheless, the tests should still provide a good indication that no order effect exists, since the observed F-values did not reach statistical significance in spite of large degrees of freedom.

However, just in case the observed heterogeneity of variance should prove to be real, two additional series of tests were carried out. The first was a series of 28 independent groups t-tests, one for each cell in the design. The t-test was selected because the individual tests would not be dependent on homogeneity of variance between cells.

For male subjects, one out of 28 cells of the design, break and enter compared to a mixed event break and enter, vandalism and assault, produced a statistically significant result ($t(17) = 2.226$, $P < 0.05$). The 28 t-tests for females subjects also produced only a single statistically significant result: break and enter compared to a mixed event break and enter and theft

($t(18) = 2.445$, $p < 0.05$). Therefore, overall, the number of significant t-tests were well within that expected by chance.

A series of Mann-Whitney U-tests were also carried out. The Mann-Whitney U-test was chosen for two principal reasons. First, the Mann-Whitney U-test is a non-parametric test and therefore does not require any assumptions regarding the distributional properties of the data. Second, the results of the U-test are invariant with respect to monotonic transformations of the data. This last characteristic can prove to be very useful if transformations of the data are to be required in model fitting. The Mann-Whitney U coefficients used in this study were computed in the normal fashion, except for ties, which were divided evenly between the two groups.

The results of the U-tests confirmed the earlier findings of both the F-test and the t-tests. None of the 28 tests for males was significant, and for female subjects, only one of 28, break and enter compared to the same mixed event break and enter and theft, was significant ($U(10,10) = 21$, $p < 0.05$). Since all three sets of tests seemed to point to a lack of an order effect, data from the two forms were combined for all subsequent analyses.

Some general observations about the data. Cell by cell inspection of the data revealed unequal cell variances in both the male and female data sets. For male subjects the smallest cell variance was 0.07 and the largest was 0.86. Levene's test for unequal variances yielded an F of 2.10. The tail probability

associated with an F of 2.1, with 27, and 493 degrees of freedom, is of approximately .0011.

The smallest variance for females was 0.05 and the largest was 0.41. Levene's test for unequal variances was also statistically significant. The computed F, with 27, and 533 degrees of freedom, was 1.93. The corresponding tail probability is .004. In both data sets, some amount of skew was also evident. The skew, together with the mean, median, standard deviation and sample size are listed in Table 1 and 2. The sign of the means and medians are arbitrary since $\ln(r(i,j)/r(i,j')) = -\ln(r(i,j')/r(i,j))$. For the sake of consistency $\bar{r}(i,j)$ was obtained from $\ln(r(i,j)/r(i,j'))$, where $j > j'$. However, with respect to a pair of descriptions, the sign of the mean or median indicate which of the two description is more serious than the other. Negative signs indicate that $r(i,j')$ is more serious than $r(i,j)$, conversely, positive signs indicate that $r(i,j)$ is more serious than $r(i,j')$.

Preliminary inspection of the data also revealed some interesting features concerning the perception of seriousness of the crime descriptions used in the study. It would appear that vandalism was considered to be more serious than theft. The average log seriousness ratio of vandalism to break and enter was 0.666 for males and 0.760 for females, while the average log seriousness ratio of theft to break and enter was approximately half as much, 0.403 and 0.305 for males and females respectively. The corresponding median log seriousness ratio was

Table 1
Mean, Median, Standard Deviation,
Sample Size and Skew
for Male Subjects

| | V | I | A | VI | VA | IA | VIA | |
|---|---|---|---|---|---|---|---|---|
| | .666 | .403 | .861 | .887 | 1.138 | 1.055 | 1.126 | Mean |
| | .620 | .405 | .908 | .776 | 1.103 | 1.081 | 1.099 | Mdn |
| B | .303 | .301 | .482 | .375 | .419 | .428 | .523 | S.D. |
| | 18 | 18 | 18 | 18 | 19 | 18 | 18 | n |
| | .310 | .015 | -.108 | .644 | .275 | 1.256 | .070 | Skew |
| | | -.589 | .798 | .199 | .981 | .595 | 1.054 | Mean |
| | | -.680 | .704 | .374 | .784 | .673 | .916 | Mdn |
| V | | .514 | .675 | .514 | .617 | .483 | .485 | S.D. |
| | | 19 | 18 | 20 | 20 | 19 | 19 | n |
| | | -.480 | .163 | -.568 | .731 | -.682 | 1.078 | Skew |
| | | | .789 | .706 | 1.017 | .967 | 1.200 | Mean |
| | | | .708 | .667 | .969 | .969 | 1.253 | Mdn |
| I | | | .257 | .438 | .547 | .410 | .418 | S.D. |
| | | | 18 | 19 | 20 | 18 | 19 | n |
| | | | .671 | .544 | .978 | .244 | -.678 | Skew |
| | | | | -.313 | .599 | .202 | .440 | Mean |
| | | | | -.405 | .405 | .139 | .146 | Mdn |
| A | | | | .927 | .638 | .270 | .343 | S.D. |
| | | | | 18 | 18 | 18 | 18 | n |
| | | | | -.129 | 1.775 | 1.433 | .292 | Skew |
| | | | | | .752 | .684 | 1.256 | Mean |
| | | | | | .696 | .452 | 1.107 | Mdn |
| VI | | | | | .382 | .534 | .636 | S.D. |
| | | | | | 20 | 19 | 19 | n |
| | | | | | .910 | 1.890 | .481 | Skew |
| | | | | | | -.158 | .249 | Mean |
| | | | | | | -.371 | .262 | Mdn |
| VA | | | | | | .593 | .309 | S.D. |
| | | | | | | 18 | 18 | n |
| | | | | | | 1.534 | -.006 | Skew |
| | | | | | | | .434 | Mean |
| | | | | | | | .405 | Mdn |
| IA | | | | | | | .458 | S.D. |
| | | | | | | | 19 | n |
| | | | | | | | .756 | Skew |

Table 2
Mean, Median, Standard Deviation
Sample Size and Skew
for Female Subjects

| | V | T | A | VT | VA | TA | VTA | |
|---|---|---|---|---|---|---|---|---|
| | .760 | .305 | .914 | .695 | 1.076 | 1.137 | 1.290 | Mean |
| | .737 | .390 | .927 | .688 | 1.952 | 1.064 | 1.139 | Mdn |
| B | .455 | .233 | .419 | .447 | .551 | .643 | .507 | S.D. |
| | 20 | 20 | 20 | 20 | 21 | 20 | 20 | n |
| | .116 | .218 | -.196 | .225 | .252 | .677 | .684 | Skew |
| | | -.573 | .624 | .220 | .918 | .993 | 1.083 | Mean |
| | | -.578 | .680 | .405 | .916 | .951 | .969 | Mdn |
| V | | .637 | .522 | .517 | .362 | .615 | .528 | S.D. |
| | | 20 | 20 | 20 | 20 | 20 | 20 | n |
| | | -.024 | -.861 | -1.122 | .563 | -.037 | .492 | Skew |
| | | | .996 | .687 | 1.075 | .867 | 1.294 | Mean |
| | | | .725 | .638 | .730 | .706 | 1.123 | Mdn |
| T | | | .615 | .425 | .630 | .510 | .476 | S.D. |
| | | | 20 | 20 | 21 | 20 | 20 | n |
| | | | .479 | .860 | .815 | 1.544 | .820 | Skew |
| | | | | -.718 | .414 | .223 | .679 | Mean |
| | | | | -.711 | .376 | .110 | .667 | Mdn |
| A | | | | .411 | .377 | .265 | .468 | S.D. |
| | | | | 20 | 19 | 20 | 20 | n |
| | | | | -.219 | .385 | .692 | .666 | Skew |
| | | | | | .934 | .652 | .675 | Mean |
| | | | | | .899 | .700 | .685 | Mdn |
| VT | | | | | .456 | .585 | .523 | S.D. |
| | | | | | 20 | 20 | 20 | n |
| | | | | | .113 | -.324 | -.570 | Skew |
| | | | | | | -.517 | .028 | Mean |
| | | | | | | -.575 | .128 | Mdn |
| VA | | | | | | .366 | .298 | S.D. |
| | | | | | | 20 | 20 | n |
| | | | | | | .538 | -1.110 | Skew |
| | | | | | | | .463 | Mean |
| | | | | | | | .497 | Mdn |
| TA | | | | | | | .391 | S.D. |
| | | | | | | | 20 | n |
| | | | | | | | -2.493 | Skew |

0.620 for males, and 0.737 for females for vandalism, and 0.405, and 0.390 for theft. This trend was also evident in mixed events. For example, vandalism in addition to a break and enter was seen to be more serious than theft in addition to a break and enter. The average log seriousness ratio of break and enter and theft to break and enter and vandalism was -0.589 for males and -.573 for females. The corresponding medians were -.68 and -.587. Negative values indicate that vandalism was seen to be more serious than theft. This trend was again evident in three-way mixed events. For example, the average log seriousness ratio of break and enter, theft and assault to break and enter, vandalism and assault was -.158 for males and -.517 for females. The medians were -.371 and -.575 respectively. Again, negative values indicate that the event which involved vandalism was more serious than the same event with theft in place of vandalism.

## Test for Sex Difference in Perceptions of Offence Seriousness

The subject of sex difference in the perception of crime seriousness has not been the focus of many studies. Sellin and Wolfgang (1964), for example, did not have much to say on the subject. Their data were collected from two groups of male students, a group of policemen and a group of juvenile court judges, who were probably all males. Akman and Normandeau (1967) and Normandeau (1966), who replicated the work of Sellin and Wolfgang, did report that some of the ratings of male and female

subjects appeared to differ from one another, but the notion of a sex difference was dismissed because the overall "shapes" of the ratings were very similar. Hsu (1973), however, did report a sex difference between the ratings of Taiwanese males and females. Hsu's conclusion was based on the the lack of similarity in "shape" and "slope", two measures which are not truly appropriate measures of mean difference.

Since the comparative ratings obtained in this study circumvent the arbitrary modulus problem, an analysis of variance similar to the one used to test for an order effect was used to test for an overall sex effect. The resulting F was statistically significant ($F(27, 1026) = 1.72$, $P < 0.05$). This finding supports the results reported by Hsu (1973).

Because of the heterogeneity of variance problem reported earlier, two additional series of test were carried out. The first was again a series of 28 independent groups t-tests. These were followed by a set of 28 Mann-Whitney U-tests.

Four of the 28 t-tests were found to be statistically significant. The comparisons which were statistically significant were, (1) vandalism compared to theft and assault, ($t(31) = -2.24$, $p < 0.05$), (2) vandalism and theft compared to vandalism, theft and assault ($t(31) = 3.12$, $p < 0.01$), (3) vandalism and assault compared to theft and assault ($t(31) = 2.27$, $p < 0.05$), and (4) vandalism and assault compared to vandalism, theft and assault, ($t(31) = 2.25$, $p < 0.05$).

U-tests produced the same four statistically significant differences between the male and female subjects. (1) vandalism compared to theft and assault ($U(19,20) = 117$, $p < 0.05$), (2) vandalism and theft compared to vandalism, theft and assault ($U(19,20) = 94$, $p < 0.05$), (3) vandalism and assault compared to theft and assault ($U(18,20) = 107$, $p < 0.05$), and (4) vandalism and assault compared to vandalism, theft and assault ($U(18,20) = 107$, $p < 0.05$). Other than the fact that all four comparisons involved assault in a mixed event, no other meaningful patterns are evident. Nevertheless, it is probably safe to assert that at least for some events, male and female subjects differed in their perception of seriousness. One of the implications of this result, with respect to the present study, is the fact that data from male and female subjects cannot be combined for purposes of analysis. Therefore, in all subsequent analyses, male and female subjects were treated separately.

## Goodness of Fit of Power Model (Females)

The result of the goodness of fit test indicated that the power model was not an appropriate model for the perceived crime seriousness data. The F-value was statistically significant ($F(21,533) = 3.56$, $p < 0.001$). This obviously has serious implications for the so-called ratio scales of offence seriousness. Wellford and Wiatrowski (1975) had hailed the development of the ratio scale of offence seriousness as "an

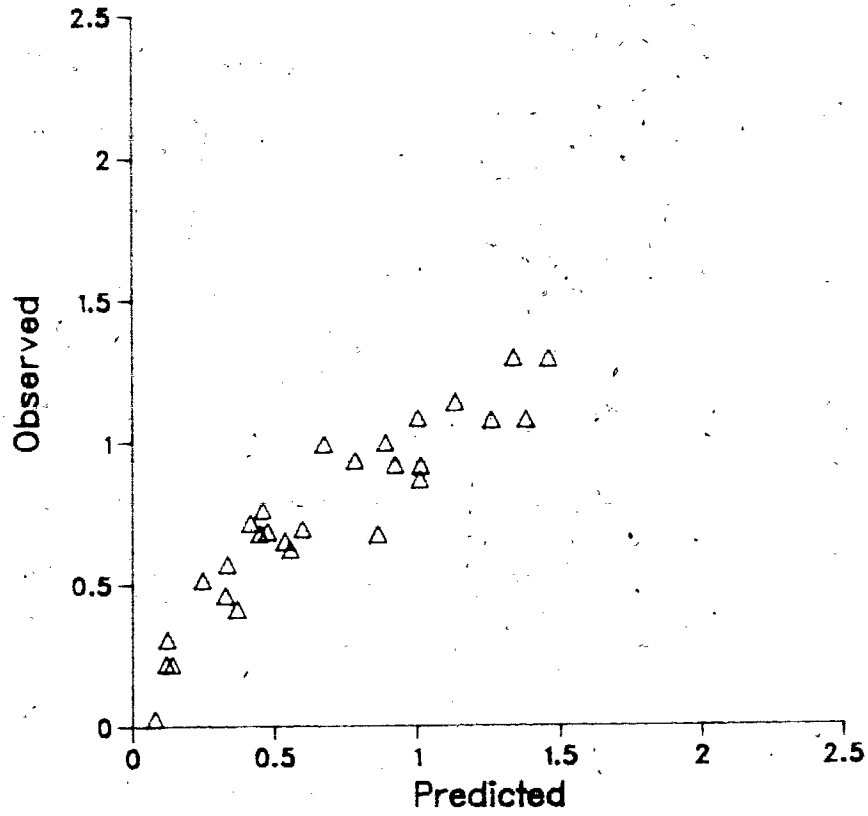
important advance in the history of criminology -- one that has provided the foundation for the development of a science of behavior" (p. 185). Can it be that this new science is built on a faulty foundation?

A plot of the observed cell means against the values predicted by the model revealed some very interesting features. Instead of the linear relationship predicted by the model, Figure 1 shows a clear curvi-linear trend. It would appear that for the more serious events, the values predicted by the model were higher than the actual observed values. Thus, in relation to the raters' perception, the model overestimated the seriousness of the more serious crimes. If this trend can be believed, the more serious the crime, the larger would be the discrepancy between observed and predicted values.

The actual lack of fit can be expected to be worse than that found in the present research, since descriptions of very serious crimes were excluded from the study. The power model failed the goodness of fit test under ideal conditions, and therefore is unlikely to fit well over the whole seriousness continuum.

Heterogeneity of variance. In defense of the power model, it must be mentioned that the results of the goodness of fit test, unfortunately, cannot be completely trusted. This is due to the heterogeneity of variance problem, discussed in an earlier section. The exact probability associated with the F-value is not known.

Figure 1
Ratio Model: Plot of Observed and Predicted
Cell means for Female subjects



## Fitting the Model with Transformed Data (females)

Researchers in psychophysics have sometimes found it necessary to correct ratings made by subjects in order to achieve linearity on the log-log plot. Ekman (1958) proposed a three parameter model which relates the magnitude estimation, R, of a rater to the physical stimulus S (equation 21),

$$(21) \qquad R = c(S - S(0))^{n}$$

where c is the usual constant related to the unit of measurement, n is the exponent which determines the amount of curvature, and S(0), the third parameter is said to represent a kind of absolute threshold.

The notion of an absolute threshold is an old one. In a review of the threshold concept, Corso (1963) stated that the term threshold was introduced into psychology by Herbart (1924). In his article, Corso (1963) distinguished the notion of a sensory threshold from a response threshold. Sensory threshold relates more to the physiological or neurological functioning of the individual, while response threshold is defined in terms of an individual's response within a specified context. The latter view permits operational definitions of the concept of threshold, and according to Corso (1963) is favored by conventional psychophysics. Regardless of whether one is

referring to sensory threshold or response threshold, one thing is clear. The perceived zero point may not necessarily be the absolute zero point, and for this reason, some researchers have found it necessary to adjust their ratings in the manner suggested by Ekman (1958). For example, McGill (1974) successfully straightened out curved loudness functions by adjusting for the absolute threshold.

Applying the concept of threshold to the scaling of psychological variables may not be as simple as it might first appear. In the scaling of physical stimuli, the numbers representing sensation are usually plotted against the physical units of the stimuli. For example in the production of loudness curves, sensation is plotted against sound intensity measured in decibels. A simple inspection of the plot can sometimes be sufficient to determine whether a threshold correction is necessary, and the approximate size of the constant, S(0), necessary to correct for the curvature can sometimes also be determined. The power model requires the plot to pass through the joint origin. If the curvature of the plot results in a line which cuts the x-axis at some point greater than zero, then the distance from zero to that point can be roughly estimated. The ratings can then be adjusted accordingly. The methods used in adjusting for an absolute threshold are often crude, but in psychophysics, they seem to work, at least in some situations.

In the scaling of psychological variables, external measurements of the variables are not generally available. The

perception of a subject cannot be plotted against an external measure to determine whether a correction for absolute threshold is necessary. Even the crude methods used in psychophysics cannot be used to adjust for an absolute threshold. However, the notion of an absolute threshold in the perception of crime seriousness seemed to be a noteworthy one, and should be pursued further.

Threshold transform. A number of attempts were made to determine the effect of the absolute threshold on the curvature of the plot of observed versus predicted cell means. Thus, a series of analyses were performed. A different threshold value was used in each analysis. The values ranged from 0.5 to 9.5, in increments of 0.5. A total of 19 analyses were therefore performed. The observed cell means were obtained from equation 22.

$$(22) \quad r(i,j,j') = \ln((r(i,j) - t) / ((r(ij') - t)$$

The term t is the threshold. The predicted cell means are obtained from a weighted least squares regression, using sample sizes as weights. A weighted regression was chosen because fitting cell means with cell sizes as weights is equivalent to fitting the raw data.

Figures 2 through 6 are plots of five of the 19 analyses. The threshold values are 1.0, 2.0, 3.0, 5.0, and 9.5, respectively. The plots clearly indicated that over the range of threshold values tested, no appreciable improvements were

Figure 2
Threshold Model: Plot of Observed and Predicted
Cell means for Female subjects
Threshold = 1.0

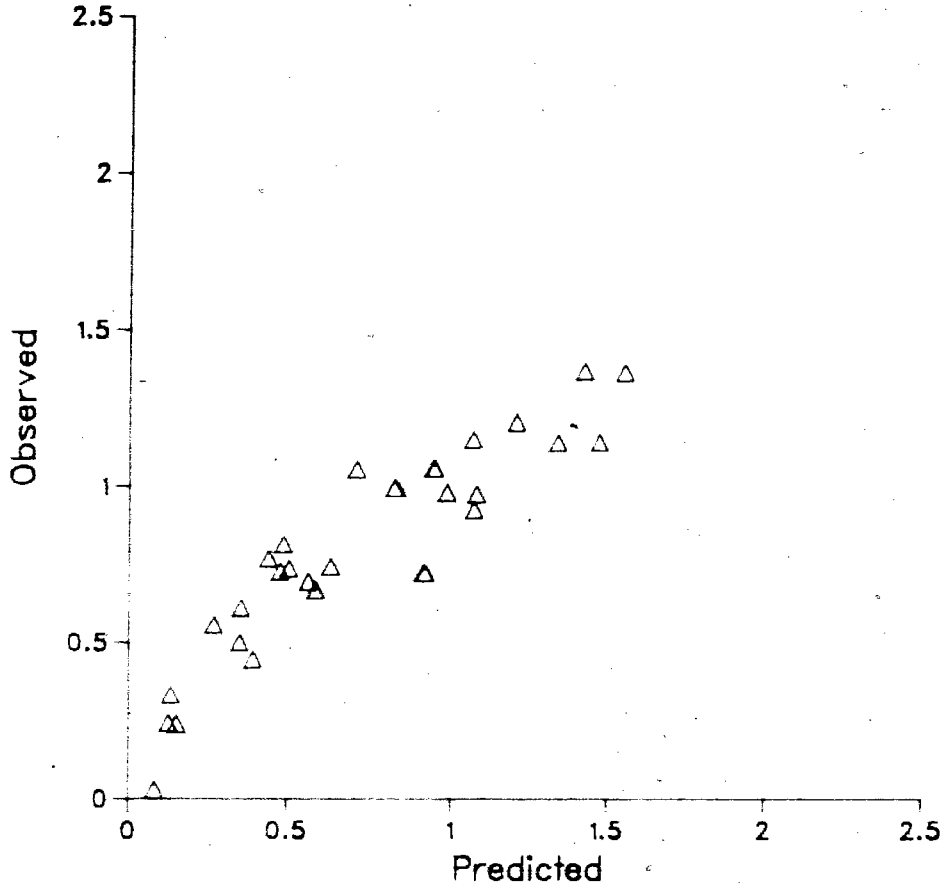


Figure 3
Threshold Model: Plot of Observed and Predicted
Cell means for Female subjects
Threshold = 2.0

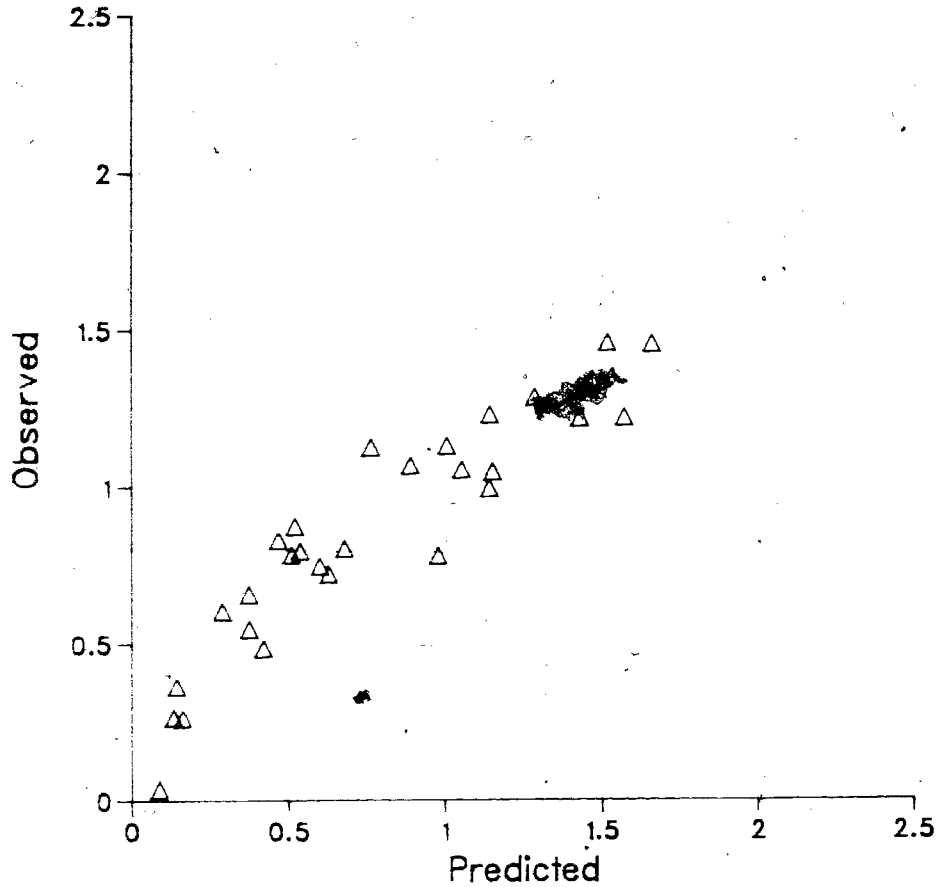


Figure 4
Threshold Model: Plot of Observed and Predicted
Cell means for Female subjects
Threshold = 3.0

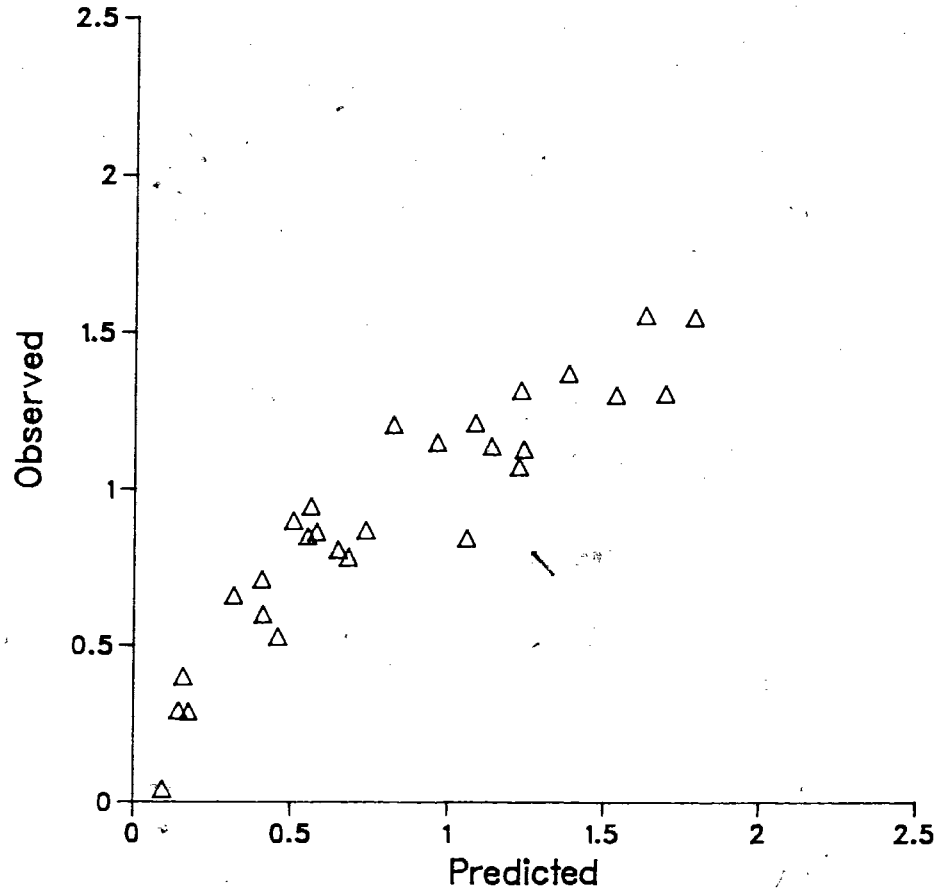


Figure 5
Threshold Model: Plot of Observed and Predicted
Cell means for Female subjects
Threshold = 5.0

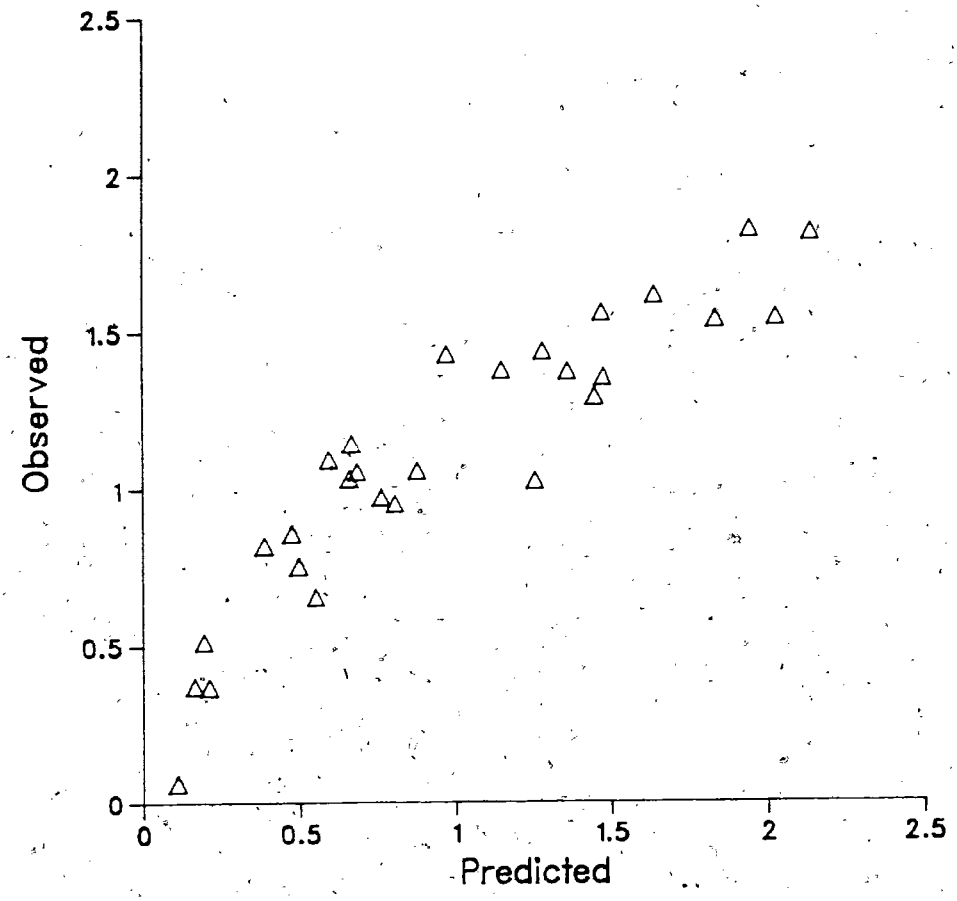
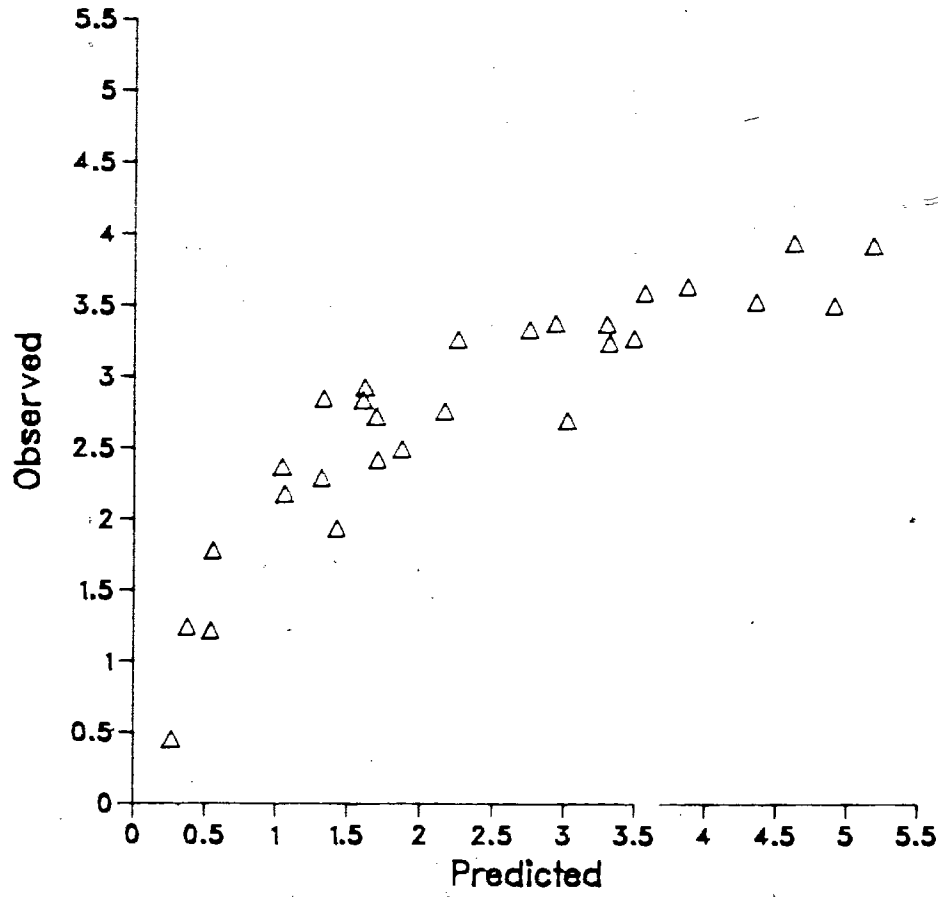


Figure 6
Threshold Model: Plot of Observed and Predicted
Cell means for Female subjects
Threshold = 9.5



obtained. This would indicate that the lack of fit in the model is probably not primarily due to an incorrect threshold and cannot be corrected by shifting the zero point of the scale to coincide with the absolute threshold.

Log-limit transform. A second series of transformations was attempted (equation 23).

$$(23)\quad r(i,j,j') = ((r(i,j)/r(i,j'))^{p} - (r(i,j')/r(i,j))^{p})/2p$$

For lack of a better name, this transformation is called the log-limit transform. The log-limit transform has some interesting properties. As p approaches zero, $r(i,j,j')$ approaches $\ln(r(i,j)/r(i,j'))$. In other words, as p approaches zero the log-limit transform approaches the ordinary log transform of the power model. As p increases in value, the convex curvature should become progressively straighter.

A total of nine p values were tested. The values of the exponent p ranged from 0.1 to 0.9 in increments of 0.1. The observed cell means were obtained from equation 23 and the predicted cell means were again obtained from a weighted least squares regression with sample sizes as weights. The plots of four of the nine transformations are displayed in Figures 7 through 10. The values of the p are 0.1, 0.4, 0.7, and 0.9 respectively.

Inspection of the plots indicated that the curvature was still quite evident. Moreover, as p increased, the scatter of the plot also increased. Clearly, the log-limit transform did

Figure 7
Loglimit Model: Plot of Observed and Predicted
Cell means for Female subjects
p = 0.1

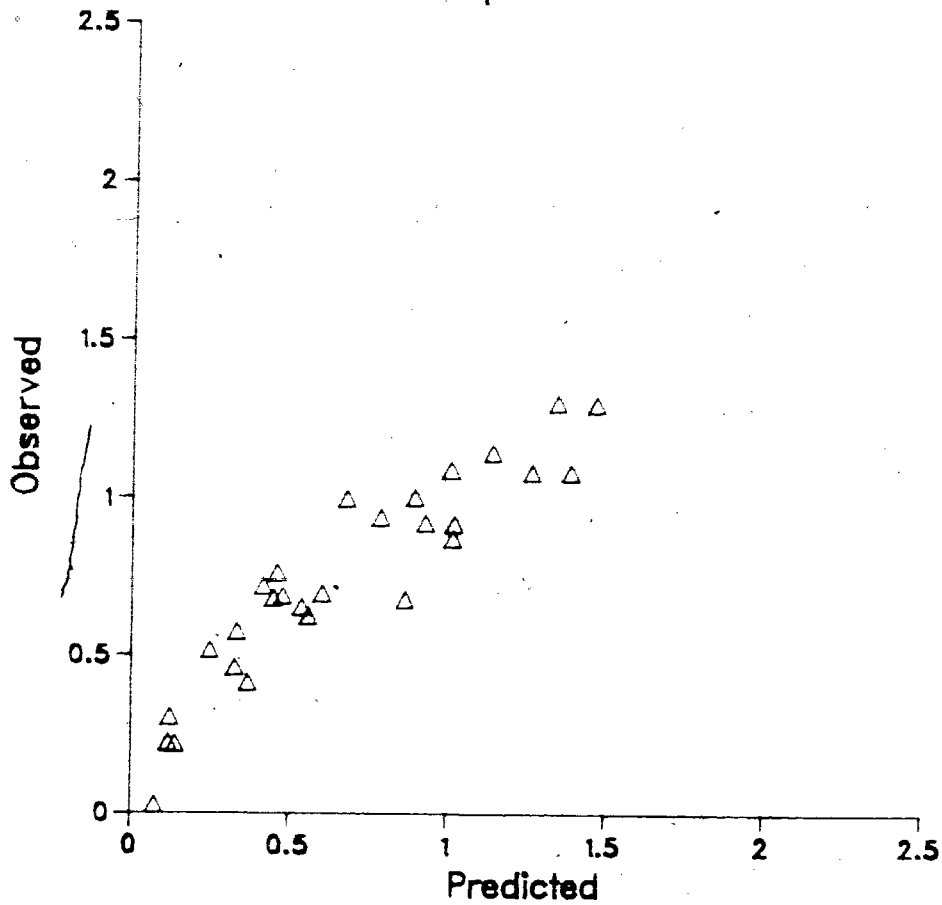


Figure 8
Loglimit Model: Plot of Observed and Predicted
Cell means for Female subjects
p = 0.4

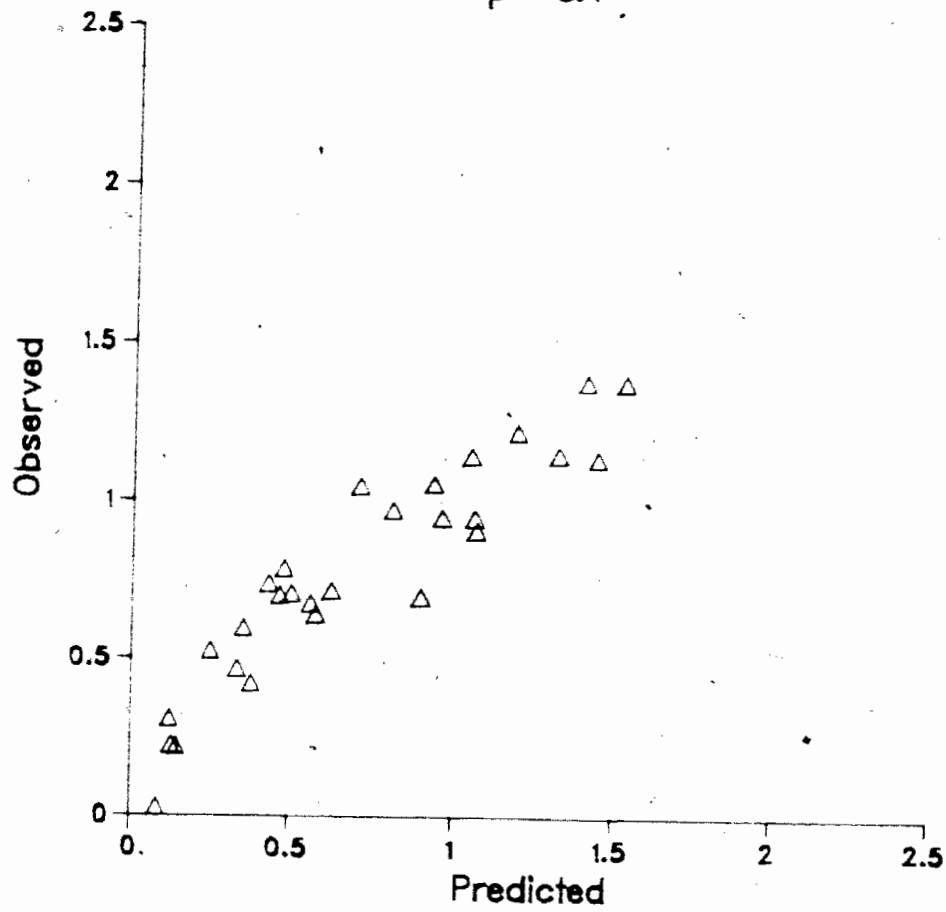


Figure 9
Loglimit Model: Plot of Observed and Predicted
Cell means for Female subjects
p = 0.7

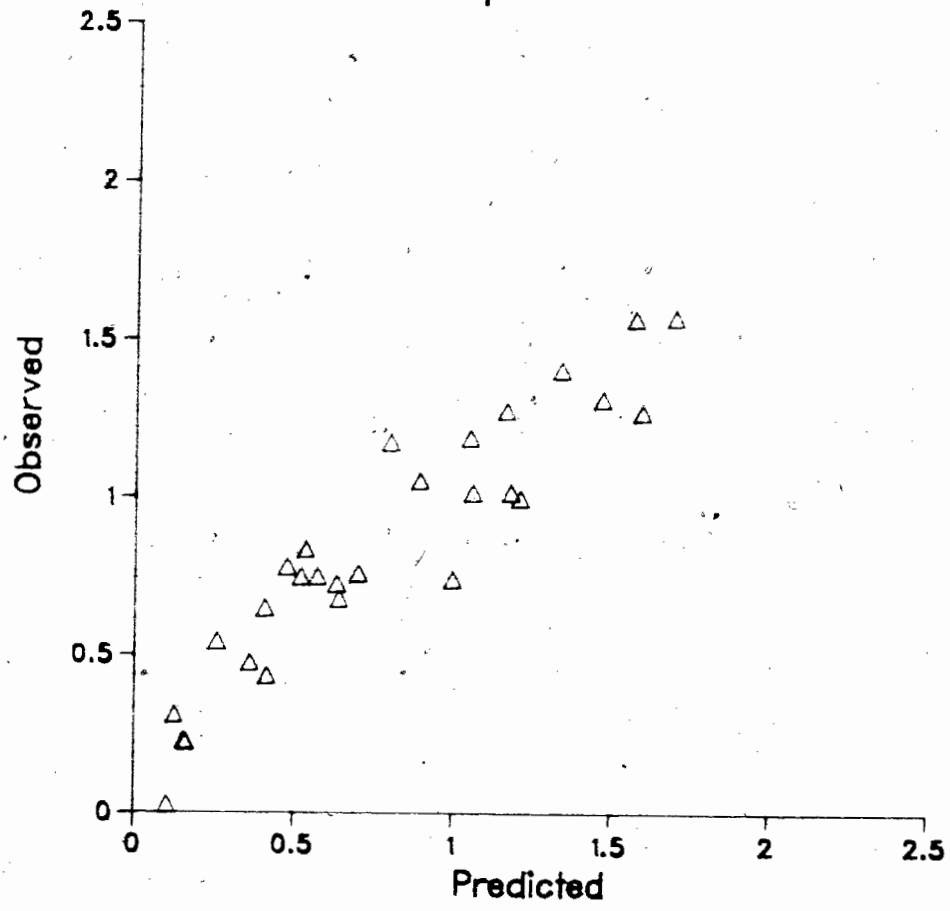
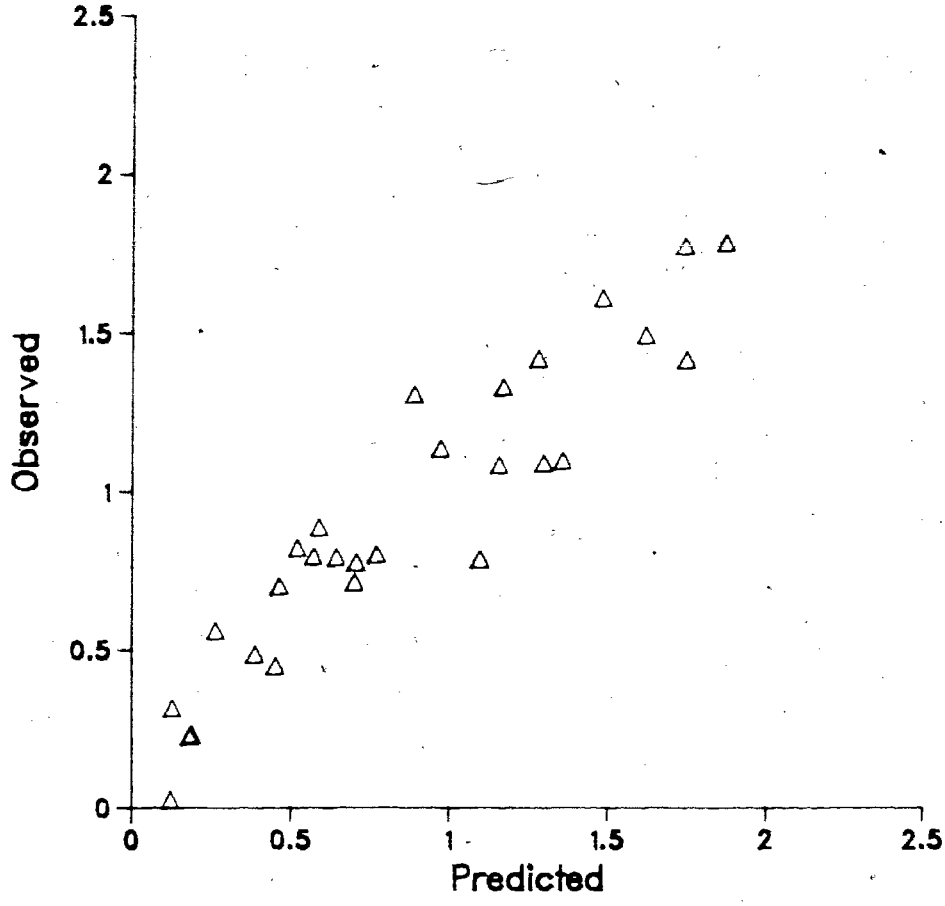


Figure 10
Loglimit Model: Plot of Observed and Predicted
Cell means for Female subjects
p = 0.9



not give a better fit in comparison to the simple log transform of the power model. The transformations were simply not powerful enough to overcome the curvi-linearity.

Quasi-linear transform. The lack of fit of the power model indicates that the relationship specified in equation 5 is not obtained. Suppose, however, that there exists some "true", but unobserved ratings t(i,j) and t(i,j'). Equation 5 can be rewritten in terms of these "true" ratings:

$$(24)\quad t(i,j)/t(i,j') = (x(j)^{b} / x(j')^{b})$$

The observed ratio of the ratings can be treated as a linear transform of the "true" ratio. That is:

$$(25)\quad r(i,j)/r(i,j') = 1 + \ln(t(i,j)/t(i,j'))$$

where $r(i,j) > r(i,j')$, and

$$(26)\quad r(i,j')/r(i,j) = 1 - \ln(t(i,j')/t(i,j))$$

where $r(i,j) < r(i,j')$. The quasi-linear transform is given by equations 27 and 28.

$$(27)\quad \ln(t(i,j)/t(i,j')) = r(i,j)/r(i,j') - 1 = r(i,j,j'),$$

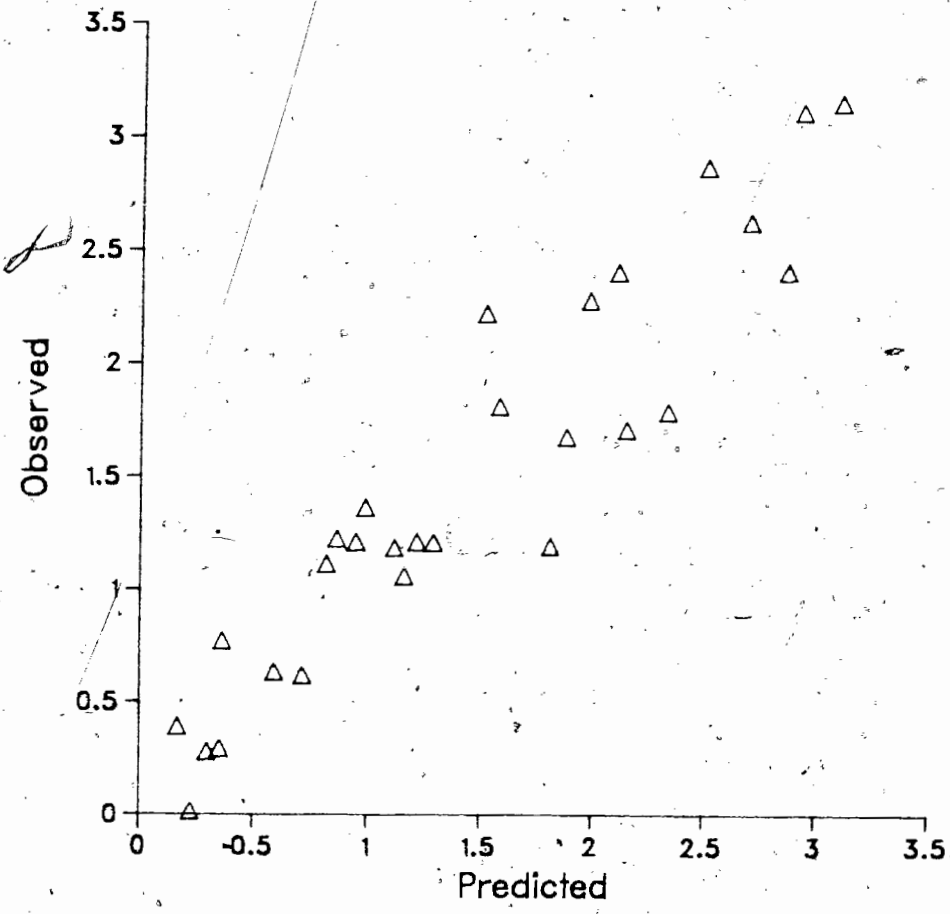
for $r(i,j) > r(i,j')$, and

$$(28)\quad \ln(t(i,j')/t(i,j) = 1 - r(i,j')/r(i,j) = r(i,j,j'),$$

where $r(i,j) < r(i,j')$.

The results of the quasi-linear transform were surprising. The plot of observed cell means against predicted cell means gave no evidence of curvi-linearity (Figure 11). Larger scores appeared to have larger variances, as the scatter in that region was more noticeable. The relationship between observed and

Figure 11
Quasi-linear Transform: Plot of Observed and Predicted Cell means for Female subjects



predicted cell means were, however, linear, as they should be. This apparent improvement in fit over the log transform of the power model was supported by the goodness of fit test ($F(21,533) = 1.06$, $P >> 0.05$). The mean, median, standard deviation, and skew are listed in Table 3. Discussion of the significance of this results is delayed until the analyses for male subjects are reported.

## Test for Additivity (females)

Power model. Sellin and Wolfgang and their supporters maintain that crime seriousness scores are additive, while their critics, notably, Pease, Ireson, and Thorpe (1974), and Rose (1966), were certain that crime seriousness scores were non-additive. The results of the present study indicated that for the types of events used in the study, additivity seemed to be confirmed. The figures in Table 4 were obtained by fitting the power model to the raw data, $r(i,j,j') = \ln(r(i,j)/r(i,j'))$. The regression weights were obtained by fitting $r(i,j,j') = CXb + e(i,j,j')$. It is evident that only main effects were statistically significant. Not one of the weights associated with an interaction term came close to reaching significance.

F-test. A reduced model comprised of the first three terms, vandalism, theft and assault, was compared to the full model with all the interaction terms included. The F for interaction was not statistically significant ($F(4,554) = 1.24$, $p > 0.05$).

Table 3
Mean, Median, Standard Deviation
Sample Size and Skew
for Female Subjects
Quasi-linear Transform

| | V | T | A | VT | VA | TA | VTA | |
|---|---|---|---|---|---|---|---|---|
| | 1.360 | .392 | 1.705 | 1.206 | 2.405 | 2.865 | 3.150 | Mean |
| | 1.050 | .481 | 1.517 | .993 | 1.688 | 1.883 | 2.150 | Mdn |
| B | 1.119 | .331 | 1.111 | 1.040 | 2.065 | 2.887 | 2.461 | S.D. |
| | 20 | 20 | 20 | 20 | 21 | 20 | 20 | n |
| | 1.200 | .602 | .742 | 1.184 | 1.712 | 1.487 | 1.589 | Skew |
| | | -1.113 | 1.060 | .280 | 1.675 | 2.220 | 2.400 | Mean |
| | | -.783 | .979 | .500 | 1.500 | 1.550 | 1.750 | Mdn |
| V | | 1.578 | 1.053 | .749 | 1.079 | 2.135 | 2.043 | S.D. |
| | | 20 | 20 | 20 | 20 | 20 | 20 | n |
| | | -1.829 | .784 | -1.154 | 1.623 | 1.767 | 1.837 | Skew |
| | | | 2.227 | 1.185 | 2.624 | 1.785 | 3.110 | Mean |
| | | | 1.050 | .940 | 1.140 | 1.020 | 2.050 | Mdn |
| T | | | 2.260 | 1.105 | 2.756 | 2.149 | 2.345 | S.D. |
| | | | 20 | 20 | 21 | 20 | 20 | n |
| | | | 1.586 | 1.770 | 1.611 | 3.237 | 1.758 | Skew |
| | | | | -1.225 | .620 | .294 | 1.210 | Mean |
| | | | | -1.050 | .467 | .117 | .950 | Mdn |
| A | | | | .980 | .640 | .370 | 1.223 | S.D. |
| | | | | 20 | 19 | 20 | 20 | n |
| | | | | -1.389 | .777 | .958 | 1.928 | Skew |
| | | | | | 1.810 | 1.210 | 1.190 | Mean |
| | | | | | 1.467 | 1.010 | .989 | Mdn |
| VT | | | | | 1.302 | 1.299 | 1.135 | S.D. |
| | | | | | 20 | 20 | 20 | n |
| | | | | | .775 | .848 | .644 | Skew |
| | | | | | | -.772 | .015 | Mean |
| | | | | | | -.783 | .015 | Mdn |
| VA | | | | | | .633 | .383 | S.D. |
| | | | | | | 20 | 20 | n |
| | | | | | | -.354 | -1.290 | Skew |
| | | | | | | | .635 | Mean |
| | | | | | | | .600 | Mdn |
| TA | | | | | | | .617 | S.D. |
| | | | | | | | 20 | n |
| | | | | | | | -2.245 | Skew |

Table 4
Regression Weights for Ratio Model
(Females)

| Effect | Regression weight | Standard error | t-statistic | 2-tail prob. |
|---|---|---|---|---|
| V | .457 | .057 | 8.05 | .000 |
| T | .123 | .057 | 2.17 | .003 |
| A | 1.012 | .057 | 17.80 | .000 |
| VT | .018 | .080 | .23 | .822 |
| VA | -.089 | .081 | -1.11 | .269 |
| TA | -.004 | .080 | -.04 | .965 |
| VTA | -.059 | .114 | -.51 | .607 |

The regression weights for the reduced model is listed in Table 5. The results of the present study therefore seemed to be consistent with the claim of additivity.

Quasi-linear Transform. The same analyses were carried out on the data with the quasi-linear transform. The regression weights are displayed in Table 6. Again, there is no evidence of interaction.

F-test. The F for interaction was also not statistically significant ($F(4,554) = 1.42$, $p > 0.05$). The weights for the reduced model are listed in Table 7. The results obtained here are similar to that obtained for the power model. Thus, the implication is clear that for female subjects, crime seriousness scales are additive, at least within the present context.

Table 5
Regression Weights for Ratio Model
Reduced (Females)

| Effect | Regression weight | Standard error | t-statistic | 2-tail prob. |
|---|---|---|---|---|
| V | .407 | .028 | 14.32 | .000 |
| T | .114 | .028 | 4.07 | .000 |
| A | .952 | .028 | 33.54 | .000 |

Table 6
Regression Weights for Quasi-linear Transform
(Females)

| Effect | Regression weight | Standard error | t-statistic | 2-tail prob. |
|---|---|---|---|---|
| V | .987 | .176 | 5.61 | .000 |
| T | .170 | .176 | .97 | .333 |
| A | 2.150 | .176 | 12.19 | .000 |
| VT | .129 | .249 | .52 | .604 |
| VA | -.270 | .250 | -1.08 | .279 |
| TA | .183 | .249 | .74 | .462 |
| VTA | -.253 | .352 | -.72 | .473 |

Table 7
Regression Weights for Quasi-linear Transform Reduced (Females)

| Effect | Regression weight | Standard error | t-statistic | 2-tail prob. |
|---|---|---|---|---|
| V | .854 | .088 | 9.68 | .000 |
| T | .263 | .088 | 2.99 | .003 |
| A | 2.043 | .088 | 23.20 | .000 |

Goodness of Fit of Power Model (males)

The goodness of fit test for male subjects also indicated that the power model was not appropriate for the data ($F(21, 493) = 3.25$, $P < 0.001$). A plot of observed cell means against predicted cell means showed a curvi-linear trend similar to that of the females (Figure 12).

Fitting the model with Transformed Data (males)

Threshold transform. In order to determine the effect of the absolute threshold on the fit of the power model for male subjects, a series of analyses, similar to the ones performed on the female data set, were carried out. The values of the threshold used ranged from 0.5 to 9.5, in increments of 0.5. Figures 13 through 17 are plots of five of the 19 analyses. The threshold values are 1.0, 2.0, 3.0, 5.0, and 9.5 respectively.

Figure 12
Ratio Model: Plot of Observed and Predicted
Cell means for Male subjects

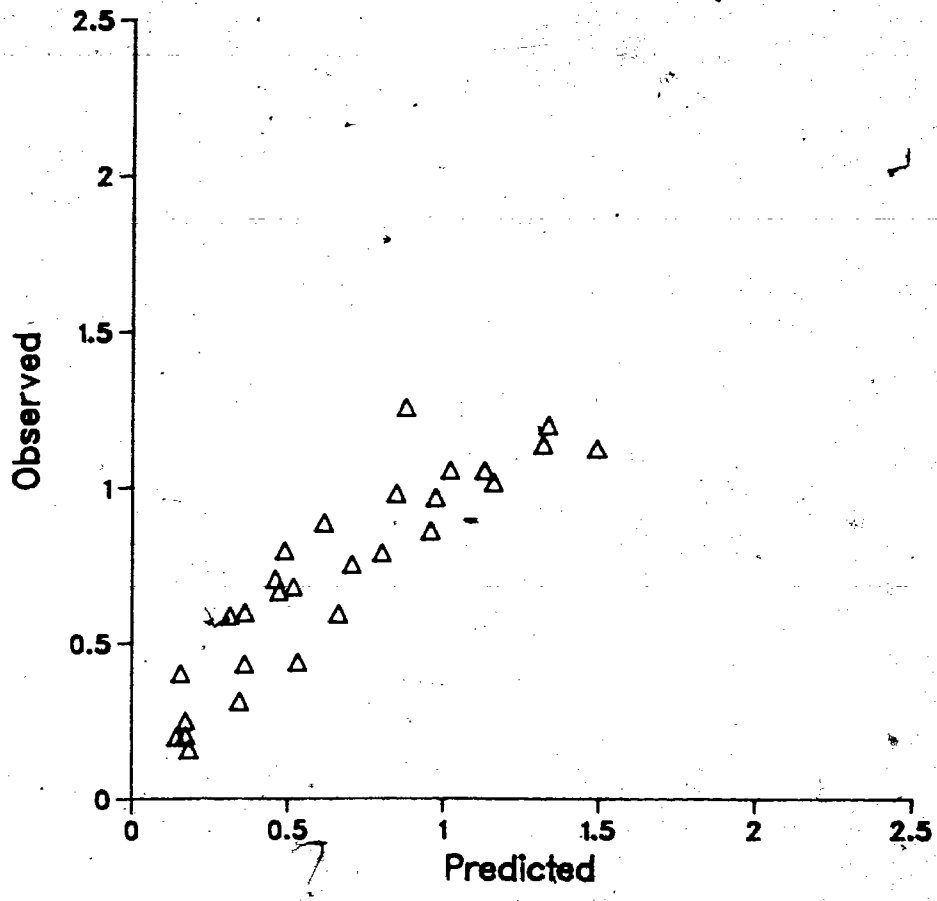


Figure 13
Threshold Model: Plot of Observed and Predicted
Cell means for Male subjects
Threshold = 1.0

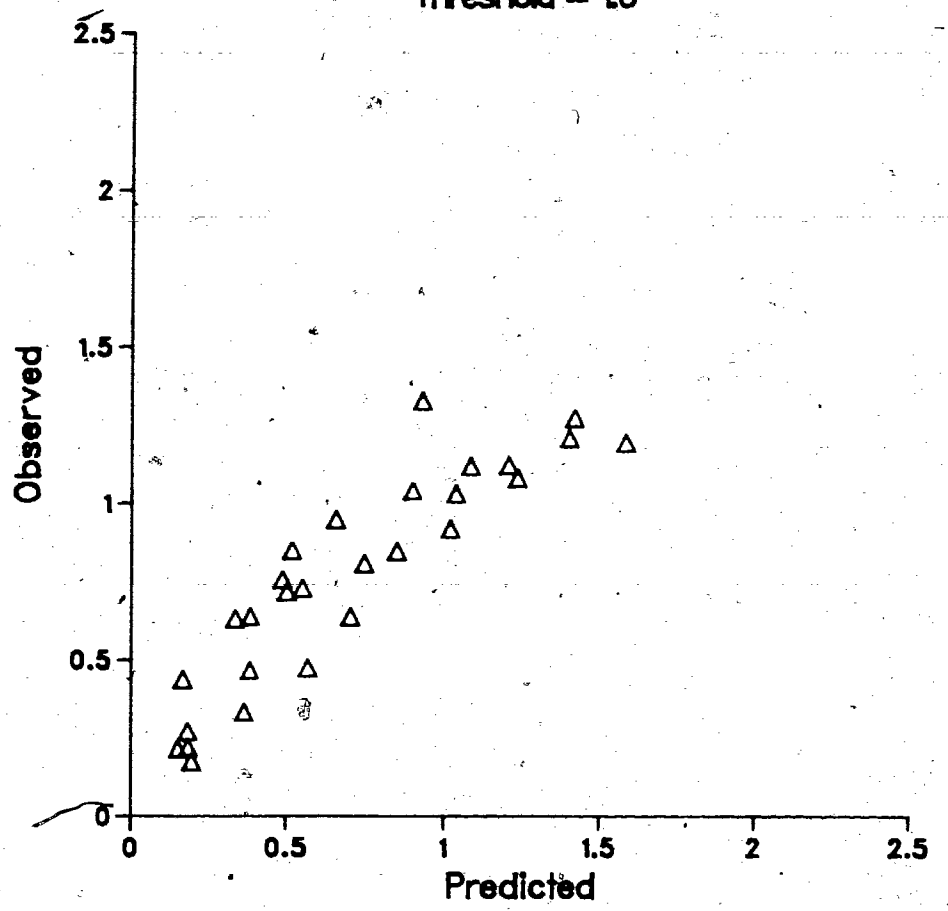


Figure 14
Threshold Model: Plot of Observed and Predicted
Cell means for Male subjects
Threshold = 2.0

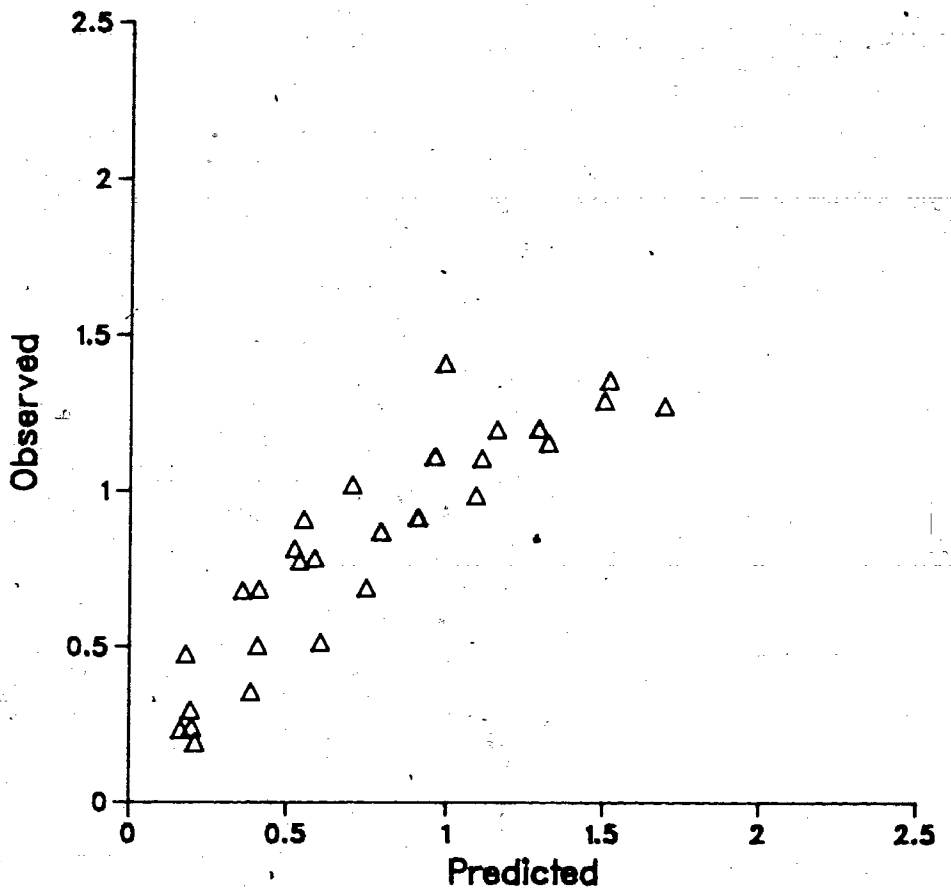


Figure 15
Threshold Model: Plot of Observed and Predicted
Cell means for Male subjects
Threshold = 3.0

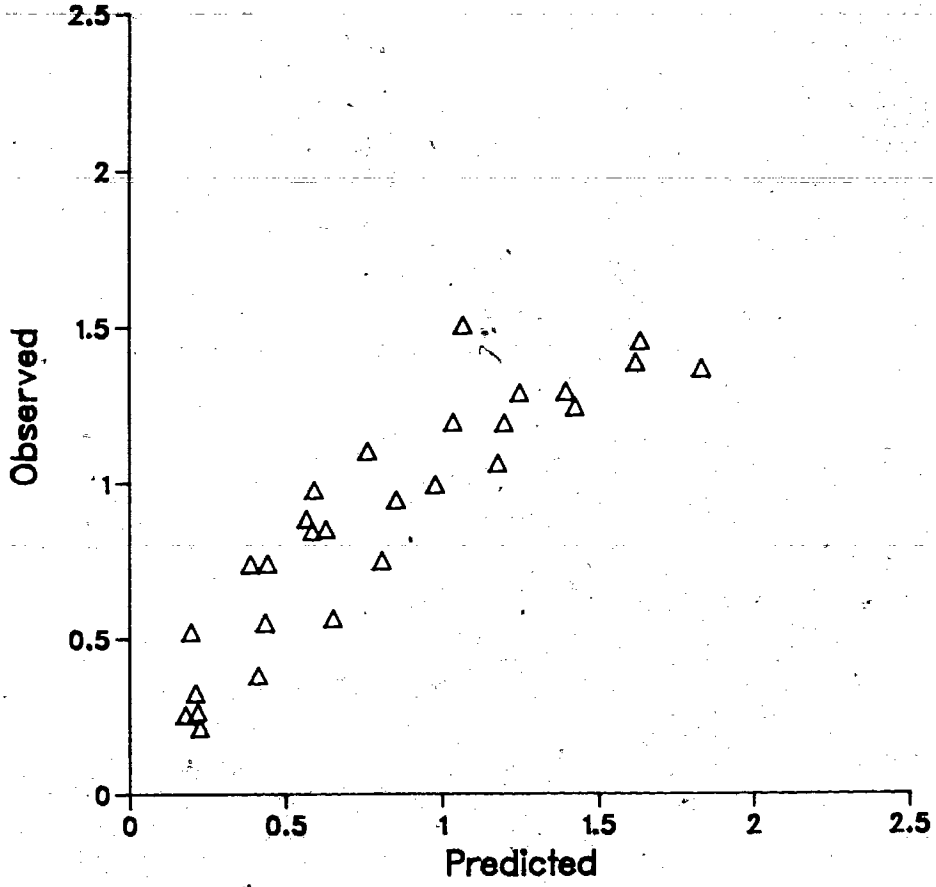


Figure 16
Threshold Model: Plot of Observed and Predicted
Cell means for Male subjects
Threshold = 5.0

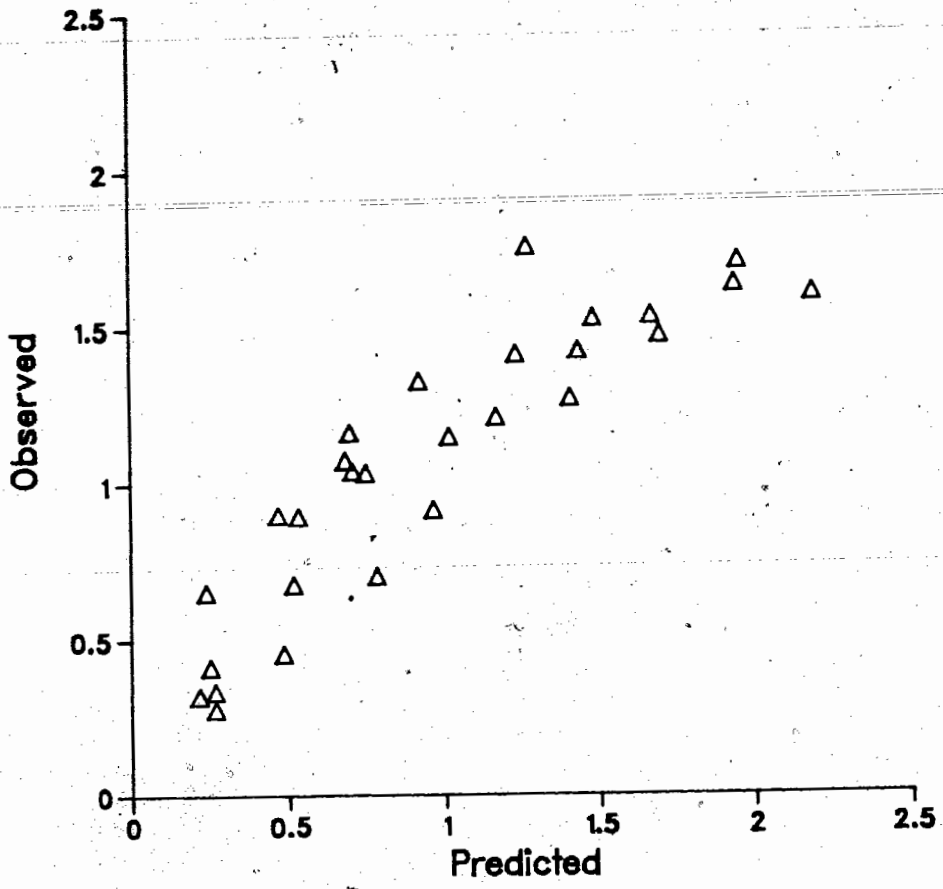
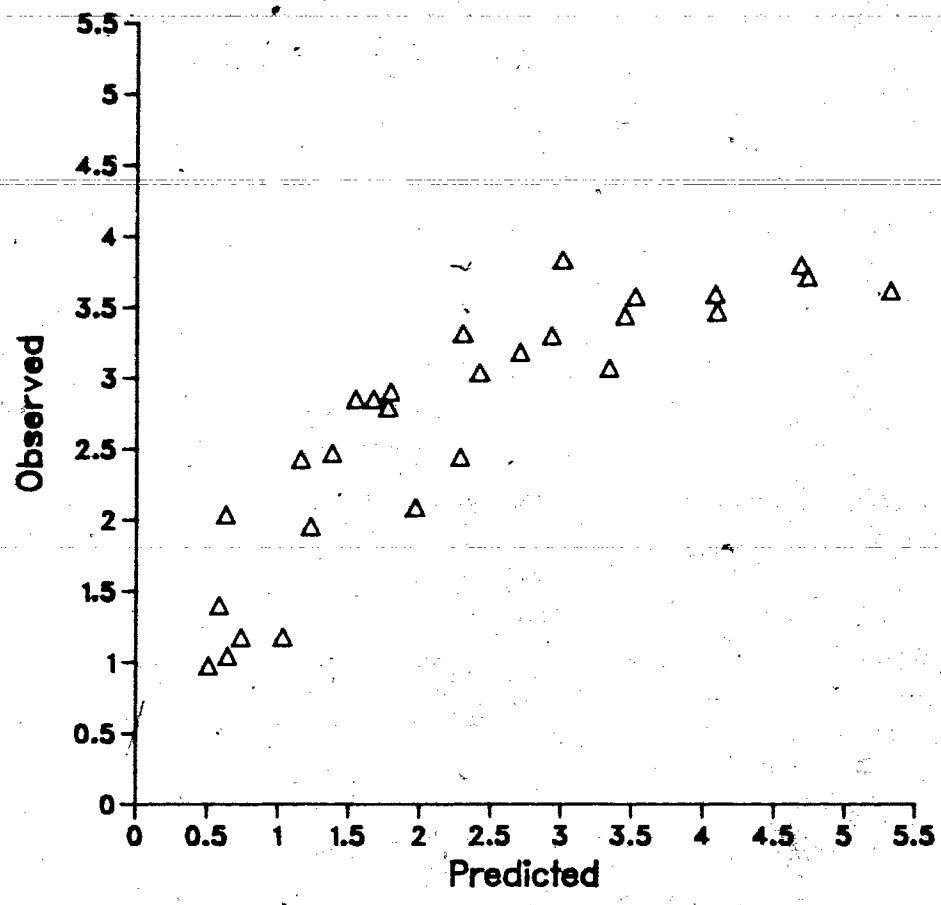


Figure 17
Threshold Model: Plot of Observed and Predicted
Cell means for Male subjects
Threshold = 9.5



Again, the plots indicated that over the range of threshold values tested, no appreciable improvements were obtained. It would appear that for both male and female subjects, an incorrect threshold is unlikely to be the major cause of the lack of fit in the model.

Log-limit transform. The log-limit transform was also used to fit the data from male subjects. The same nine values of the exponent used for females were used here for male subjects. The plots of four of the nine transformations are displayed in Figures 18 through 21.

Inspection of the plots indicated that the curvature was still present. The use of the log-limit transform, therefore, did not result in a better fit for either male or female subjects.

Quasi-linear transform. Since the quasi-linear transform seemed to have produced a better fit for female subjects, it was hoped that fitting the model to data from males would yield the same result. The mean, median, standard deviation, and skew are listed in Table 8. Unfortunately, the goodness of fit test was statistically significant ($F(21,493) = 1.67$, $p < .05$). However, an F of 1.67 with 21 and 493 degrees of freedom is just barely significant at the 0.05 level. This indicated that the quasi-linear transform did produce a somewhat better fit than the power model. The improvement is also evident by simply looking the plot of observed versus predicted cell means (Figure 22). No curvi-linearity is evident.

Figure 18
Loglimit Model: Plot of Observed and Predicted
Cell means for Male subjects
p = 0.1

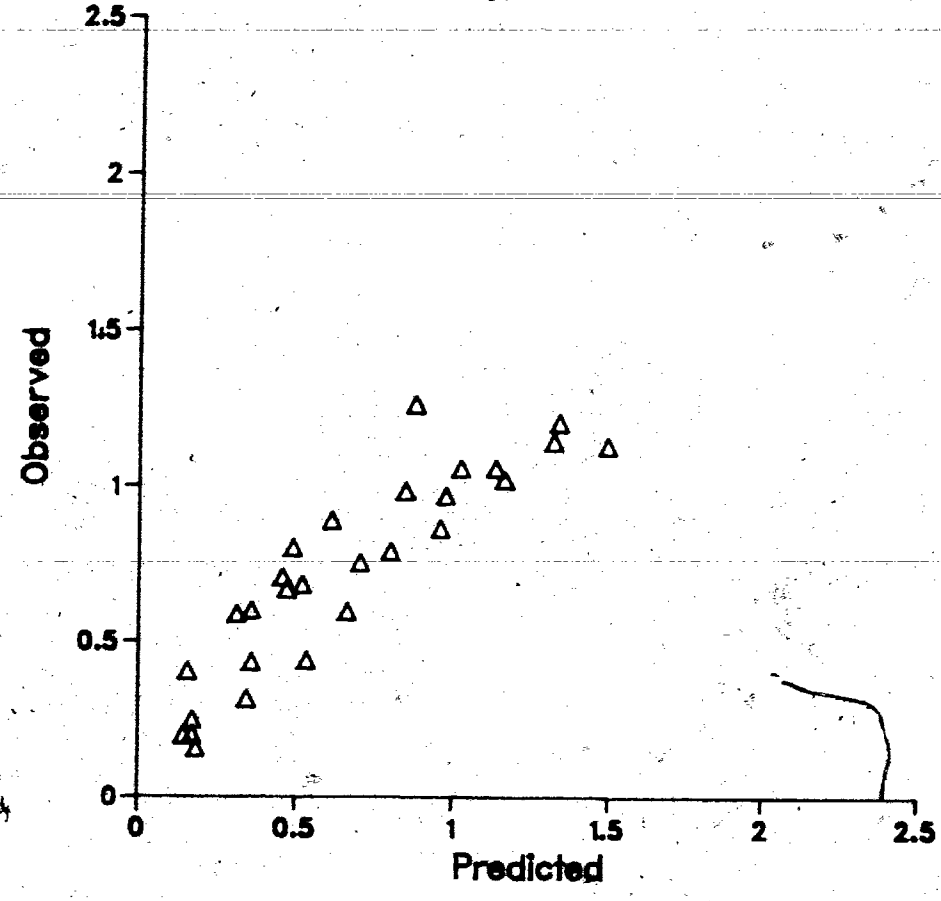


Figure 19
Loglimit Model: Plot of Observed and Predicted
Cell means for Male subjects
p = 0.4

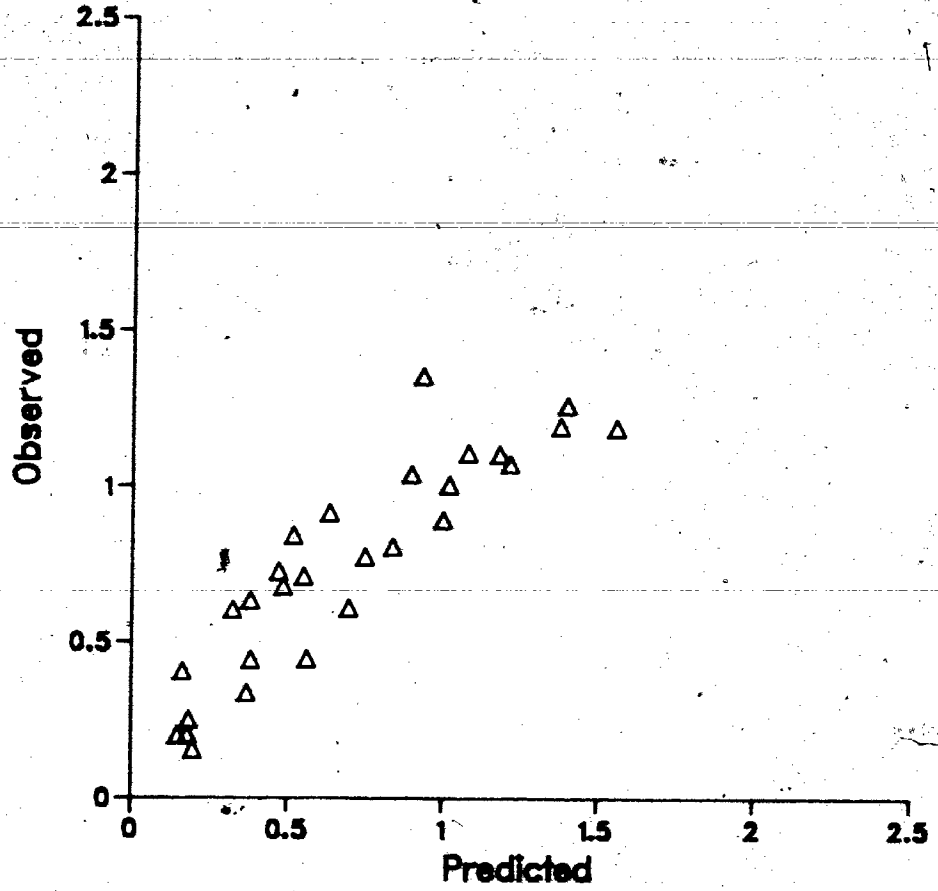


Figure 20
Loglimit Model: Plot of Observed and Predicted
Cell means for Male subjects
p = 0.7

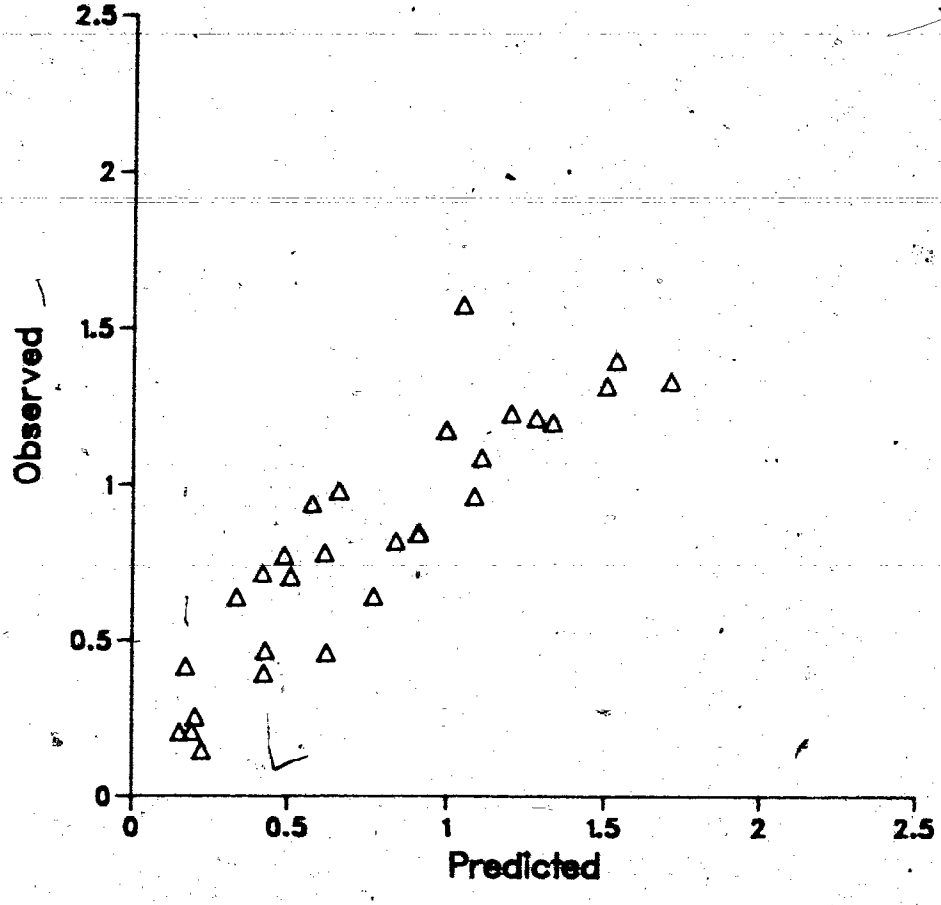


Figure 21
Loglimit Model: Plot of Observed and Predicted
Cell means for Male subjects
p = 0.9

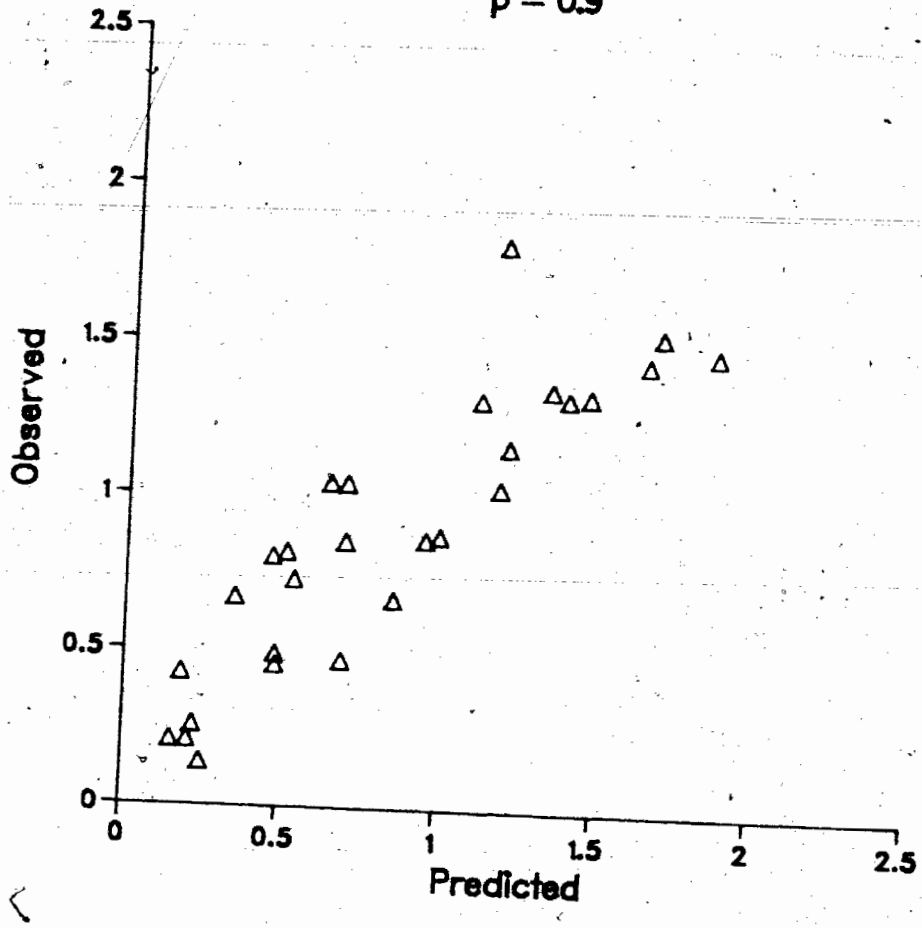


Figure 22
Quasi-linear Transform: Plot of Observed and Predicted
Cell means for Male subjects

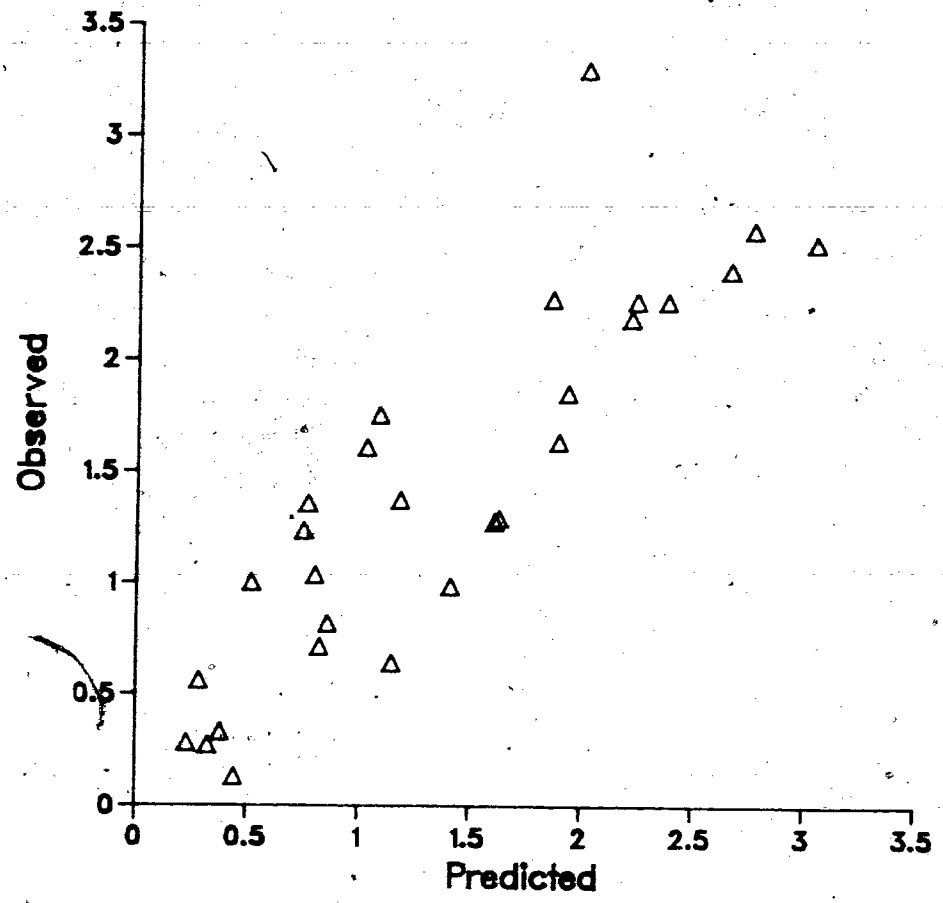


Table 8
Mean, Median, Standard Deviation,
Sample Size and Skew
for Male Subjects
Quasi-linear Transform

| | V | T | A | VT | VA | TA | VTA | |
|---|---|---|---|---|---|---|---|---|
| | 1.033 | .561 | 1.633 | 1.606 | 2.400 | 2.183 | 2.519 | Mean |
| | .850 | .500 | 1.483 | 1.150 | 2.006 | 1.967 | 2.000 | Mdn |
| B | .633 | .465 | 1.261 | 1.083 | 1.535 | 1.872 | 2.046 | S.D. |
| | 18 | 18 | 18 | 18 | 19 | 18 | 18 | n |
| | .582 | .337 | .896 | 1.289 | 1.350 | 3.066 | 2.008 | Skew |
| | | -1.001 | 1.752 | .280 | 2.270 | .982 | 2.263 | Mean |
| | | -.982 | 1.017 | .467 | 1.100 | .969 | 1.500 | Mdn |
| V | | 1.063 | 2.237 | .786 | 2.529 | 1.003 | 2.030 | S.D. |
| | | 19 | 18 | 20 | 20 | 19 | 19 | n |
| | | -.986 | 2.184 | -.261 | 2.065 | 1.210 | 2.305 | Skew |
| | | | 1.274 | 1.232 | 2.261 | 1.852 | 2.579 | Mean |
| | | | 1.036 | .971 | 1.600 | 1.667 | 2.500 | Mdn |
| T | | | .634 | 1.087 | 2.320 | 1.206 | 1.305 | S.D. |
| | | | 18 | 19 | 20 | 18 | 19 | n |
| | | | 1.304 | 1.329 | 2.193 | .786 | -.214 | Skew |
| | | | | -.817 | 1.357 | .272 | .643 | Mean |
| | | | | -.500 | .500 | .150 | .517 | Mdn |
| A | | | | 2.537 | 2.336 | .410 | .586 | S.D. |
| | | | | 18 | 18 | 18 | 18 | n |
| | | | | -2.065 | 2.587 | 1.998 | .870 | Skew |
| | | | | | 1.290 | 1.371 | 3.295 | Mean |
| | | | | | 1.007 | .587 | 2.012 | Mdn |
| VT | | | | | 1.044 | 2.055 | 2.959 | S.D. |
| | | | | | 20 | 19 | 19 | n |
| | | | | | 1.870 | 3.243 | 1.152 | Skew |
| | | | | | | -.131 | .331 | Mean |
| | | | | | | -.450 | .300 | Mdn |
| VA | | | | | | 1.193 | .445 | S.D. |
| | | | | | | 18 | 18 | n |
| | | | | | | 2.459 | .813 | Skew |
| | | | | | | | .716 | Mean |
| | | | | | | | .500 | Mdn |
| TA | | | | | | | .985 | S.D. |
| | | | | | | | 19 | n |
| | | | | | | | 2.250 | Skew |

Table 9
Regression Weights for Ratio Model
(Males)

| Effect | Regression weight | Standard error | t-statistic | 2-tail prob. |
|---|---|---|---|---|
| V | .469 | .060 | 7.78 | .000 |
| T | .157 | .060 | 2.60 | .010 |
| A | .956 | .061 | 15.69 | .000 |
| VT | -.013 | .085 | -.16 | .877 |
| VA | -.01 | .085 | -1.29 | .197 |
| TA | .017 | .086 | .19 | .846 |
| VTA | .013 | .120 | .11 | .912 |

Test for Additivity (Males)

Power model. The results of tests for additivity were consistent with those obtained with female subjects. It is quite clear that only main effects were statistically significant. None of the weights associated with the interaction terms reached statistical significance (Table 9).

A comparison of the reduced power model (Table 10) with the full power model (Table 9) also indicated that the four interaction terms, taken as a whole, were not statistically significant ($F(4,514) = .786$, $p > 0.05$). These results were again consistent with the claim of additivity.

Analogous results were obtained from data with the quasi-linear transform. None of the interactions terms was statistically significant. The comparison of the reduced model with the full model (Tables 11, 12) produced an F for

Table 10
Regression Weights for Ratio Model
Reduced (Males)

| Effect | Regression weight | Standard error | t-statistic | 2-tail prob. |
|---|---|---|---|---|
| V | .411 | .030 | 13.67 | .000 |
| T | .162 | .030 | 5.38 | .000 |
| A | .911 | .030 | 30.50 | .000 |

Table 11
Regression Weights for Quasi-linear Transform
(Males)

| Effect | Regression weight | Standard error | t-statistic | 2-tail prob. |
|---|---|---|---|---|
| V | .800 | .188 | 4.25 | .000 |
| T | .281 | .189 | 1.49 | .136 |
| A | 1.890 | .190 | 9.93 | .000 |
| VT | -.051 | .264 | -.19 | .848 |
| VA | -.031 | .266 | -.12 | .906 |
| TA | .044 | .268 | .16 | .871 |
| VTA | .106 | .375 | .28 | .778 |

Table 12
Regression Weights for Quasi-linear Transform
Reduced (Males)

| Effect | Regression weight | Standard error | t-statistic | 2-tail prob. |
|---|---|---|---|---|
| V | .786 | .094 | 8.40 | .000 |
| T | .304 | .094 | 3.25 | .000 |
| A | 1.923 | .093 | 20.65 | .000 |

interaction which was not statistically significant, ($F(4,514) = .09$, $p > 0.05$). Taken as a whole, these results indicate that

Sellin and Wolfgang may indeed have been correct when they assumed that crime seriousness scores were additive.

## VI. Conclusions

Since the results of the present study are rather complex, the conclusions which can be drawn from them are also likely to be something less than perspicuous. The primary factor responsible for this state of affairs is heterogeneity of variance, the feared characteristic which real data sometimes tend to have. In the present study, the violation of the homogeneity of variance assumption meant that the results of many of the critical tests were difficult to assess, because the exact probabilities associated with the tests are not known. Nevertheless, some important patterns did emerge.

Sex difference. The results of the statistical tests were consistent. There appears to be little doubt that for some events, the perception of males and females differ. From a practical point of view the discrepancies may not be too serious. However, it is reasonable to speculate that for sexually charged crimes such as sexual assault and rape, events which were purposely excluded from this study, the difference will be somewhat more pronounced. This finding has some implications for offence seriousness scales which do not recognize the divergent views of males and females. It is not clear at this time how a sex difference can be incorporated in a measure of offence seriousness, however, it is clear that some re-thinking is now in order.

Fit of the power model. The most basic issue addressed by this study is the question as to whether or not the power model is an appropriate one for offence seriousness data. Because of the complexity of the results, the easiest position to take on this issue is to take no position at all. But if one must take a position, the answer has to be a qualified no. From one point of view, the performance of the power model is quite respectable. It accounted for some 71 per cent of the variance in the female data set and 69 per cent of the variance in the male data set. However, there is a systematic error in the fit of the model. Because of the curvi-linear relationship, the fit is acceptable for moderately serious crime, but it is expected to become progressively worse for the more serious crimes. This error was evident for both male and female subjects.

The negative assessment of the fit of the power model is supported by the results of the goodness of fit test. For both male and female subjects, test results were highly significant. The difficulty here, however, is one of interpretation. Since the homogeneity of variance assumption was violated, the exact probabilities associated with the F-values are not known. Nevertheless, taken as a whole, the pattern is difficult to overlook.

Fit of the model with transformed data. Neither the threshold nor the log-limit transformations resulted in any appreciable improvements. In addition to the weighted least squares regressions reported earlier, an additional series if

regressions was performed using $n(j,j')/V(j,j')$, sample size and the reciprocal of the cell variance as weights, instead of sample size alone. The resulting plots of the observed and the predicted means also exhibited a definite curvi-linear trend, indicating that no advantage was gained by paying less attention to cells with larger variances.

Fit of the model with quasi-linear transform. The quasi-linear transform produced some very interesting results. The model did not account for as much of the variance in the data as the simple power model. For male subjects, the quasi-linear transform produced a fit which accounted for 50 per cent of the variance in the data. For female subjects, the corresponding amount was 54 per cent. The fit, however, was linear in both subject groups. The goodness of fit tests were also much improved. The test results were not statistically significant for female subjects, and barely made it at the 5 per cent level of confidence for male subjects. In this respect, the quasi-linear transform produced a substantial improvement in fit over the power model. Squared multiple correlations of 0.50 for males and 0.54 for females might appear to be rather low, given the fact that correlations in the 0.90s have been frequently reported in this area of research. However, in general, these correlations are often obtained by fitting cell means, not by fitting the data, as is the case in the present study. In fact, with the quasi-linear transform, the correlations between observed and predicted cell means were 0.97 for males and 0.98

for females, and the corresponding squared correlations were 0.94 for males and 0.97 for females. From this point of view, the fit is not at all bad.

The heterogeneity of variance evident in the power model was more pronounced in the case of the quasi-linear transform. The log transform of the power model kept the larger variances under "check", but no analogous rescaling was performed on the data in the present case. As a result, the larger variances remained large. This is evident by inspecting Figures 11 and 22, where a fan-like pattern can be observed. Nevertheless, the linear trend produced by the quasi-linear transform, and the better results on the goodness of fit test indicate that the use of the power model in this context should be carefully reconsidered.

Interaction. The views of Sellin and Wolfgang (1964) and their followers regarding additivity seemed to have been confirmed by the present research. But here again, a few statements of qualification are necessary. While interaction was found not to be a factor in both the ratio and the interval model for both male and female subjects, the problem is, again, one of interpretation. How meaningful are the results when the model is known not to fit to begin with? In the case of the power model there is a real likelihood that the model is wrong. In order to bias the results in favor of the power model by selecting crime descriptions of moderate seriousness, it is conceivable that perceived offense seriousness may have been

restricted to a region where the subjects' responses are additive. It is, of course quite possible that the same results might not generalize to other parts of the seriousness continuum. In this respect, the two goals, goodness of fit and interaction are, not completely compatible with one another. Thus, strictly speaking the finding of additivity should not be extrapolated beyond the range of seriousness used in this study. However, given the fact that the results were consistent across both sexes, and models, one can perhaps afford a certain measure of optimism and be allowed to join the ranks of the supporters of Sellin and Wolfgang, at least in regard to additivity of crime types.

An aspect of additivity not handled by this study is additivity with respect to multiple incidents of the same crime, or some mixture of multiple incidents of the same crime and other crimes. Until all possible aspects have been investigated, additivity remains, for the moment, an open issue.

Some closing comments. The present study was carried out in an attempt to validate Sellin and Wolfgang's approach to the scaling of offence seriousness, and to clarify a number of outstanding issues, namely sex difference and additivity of crime types. Having performed the study, examined the data, and reflected upon the results, one can conclude that although the results showed that the fit of the power model is not good, for the present, the power model should not be discarded, without further attempts to uncover and understand the factors

responsible for its lack of fit. There is, however, an inadequate information base within the area of offence seriousness scaling to guide further speculations concerning the results obtained in the present study. Until an adequate data base is built up, one must look elsewhere.

Criminologists have too often overlooked the potential relevance of research in psychophysics. For example, it has been known for some time that the size of the modulus has an effect on magnitude estimations. Typically, the size of the exponent is inversely related to the size of the modulus (Engen & Ross, 1966; Wong, 1963). Thus, by simply varying the size of the modulus, experimenters can induce a change in the observed size of the psychophysical exponent. Whether this relationship applies to the scaling of offence seriousness is not known, but inspection of the data revealed that the modulus did have an observable effect. Almost all of the ratings were either multiples of 10 or 5. This is perhaps not too surprising, since the modulus used in the experiment is 10. While the model specifies a continuous scale, the subjects responded with discrete numbers. This inability of subjects to respond in the expected manner may be a reflection of a distortion of the subjective seriousness scale, brought about by the experimental procedure. The effect of the distortion caused by this phenomenon on the fit of the model is, of course, undetermined. Nevertheless, as a source of error, the effect of the modulus should be minimized whenever possible.

The literature in psychophysics refer to a regression effect which Stevens and Greenbaum (1966) described as a tendency on the part of the rater to constrict or shorten the range of the variable under adjustment. In other words, the subject has a tendency to avoid the use of large numbers, and thereby "regress" towards the mean. Whether or not the regression phenomenon operates in judgments of offence seriousness is uncertain, but if there is regression, and if for some reason the regression is non-linear, it could, in part, explain the results obtained in this study.

The application of a model to a new field without adequate groundwork is inadvisable. One must first thoroughly understand the basic relationships, and limitations. Without this knowledge, effects cannot be distinguished from artifacts. There is much more to magnitude estimation than the mere use of magnitude scaling techniques. Thus, with respect to offence seriousness scales, basic research should be given first priority.

The possible existence of a sex difference underlines the importance of determining the perceptual and theoretical parameters of the term "offense seriousness". If the perception of offense seriousness can be influenced by such things as intent, sex and age of the victim, punitiveness of the rater, then the degree to which these variables affect seriousness ratings must be empirically determined. There is some evidence that factors other than those indicated by Sellin and Wolfgang

(1964), such as amount of injury, theft or damage play a part in the determination of seriousness. McCleary, O'Neil, Epperlein, Jones, and Gray (1980) for example, have shown that such variables as formal legal education and work experience underlie consensus. They further found that while the ordinary citizen perceives the relative seriousness of a crime in terms of only a few dimensions, criminal justice workers perceive seriousness in terms of many dimensions. It is reasonable to speculate that the perception of offense seriousness is a multi-dimensional phenomenon, and researchers are only just beginning to realize the complexity of the problem.

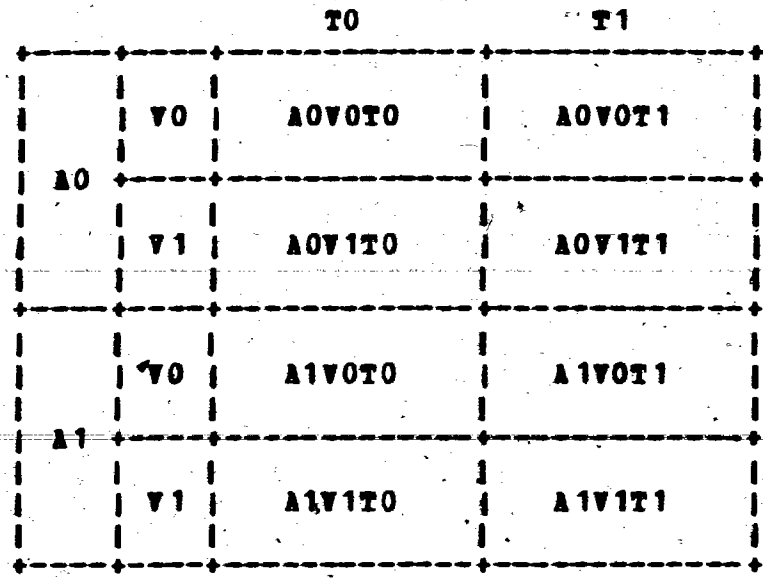
Some ideas for future research. In the study of the perception of offence seriousness, the individual is the measuring instrument. Reason would suggest that in order to understand the perception of offence seriousness, one needs to delineate the operational characteristics of the measuring instrument. In this respect, it would be interesting to determine whether or not the transition from perception to the production of a response is affected by the nature of the required response. In this study, for example, the responses produced by the subjects were clearly related to the modulus used. However, the fact that the measured "difference" came in multiples of ten or five does not necessarily mean that the subjects' actual perceived "difference" is in multiples of ten or five. This may be result of the requirement that subjects translate their perception to numbers. Information may be lost

or modified in the process.

A smoother set of responses can perhaps be obtained by asking subject to, say, adjust the length of two lines or the areas of two circles. With the advent of micro-computers with built in graphics and interfaces to measuring instruments, this can be accomplished with relative ease. Thus, the effect of types of responses on the perception of offence seriousness can be studied. And, by varying the modulus, the effect of the modulus can also be studied. It is known, for example, that the slope of the power model is related to the modulus (Ross & Engen, 1966)

There are no rules which require that subjects respond with numbers in a magnitude estimation task. A verbal response is just one way of getting at the relationships inside the "instrument". There may be other methods which are more reliable and less susceptible to procedure imposed artifacts. Moreover, the search for alternate methods may well lead to the discovery of some common patterns, and thus lead to a better understanding of how one evaluates the relative seriousness of offenses.

# APPENDIX A




A = Assault
V = Vandalism
T = Theft

## APPENDIX B

A0T0V0
An intruder broke into a ground floor apartment by forcing open a locked door. He looked around. As he thought he heard someone coming home, he fled through a window without being seen.

A0T0V1
An intruder broke into a ground floor apartment by forcing open a locked door. He looked around, then slashed the paintings on the wall, and smashed the mirror in the hallway. As he thought he heard someone coming home, he fled through a window without being seen.

A0T1V0
An intruder broke into a ground floor apartment by forcing open a locked door. He looked around, then took the colour television set. The intruder fled through a window without being seen.

A1T0V0
An intruder broke into a ground floor apartment by forcing open a locked door. As he was looking around, the tenant returned and surprised him. The intruder struck the tenant and fled. The tenant was hurt, and he was taken to hospital.

A0T1V1
An intruder broke into a ground floor apartment by forcing open a locked door. He looked around, then took the colour television set. On his way out, he slashed the paintings on the wall, and smashed the mirror in the hallway. The intruder fled through a window without being seen.

A1T0V1
An intruder broke into a ground floor apartment by forcing open a locked door. He slashed the paintings on the wall, and smashed the mirror in the hallway. The tenant returned and surprised him. The intruder struck the tenant and fled. The tenant was hurt, and he was taken to hospital.

A1T1V0
An intruder broke into a ground floor apartment by forcing open a locked door. He looked around, then took the colour television set. Just as he was about to leave, the tenant returned and surprised him. The intruder struck the tenant and fled with the television set. The tenant was hurt, and he was taken to hospital.

A1T.1V1
An intruder broke into a ground floor apartment by forcing open a locked door. He looked around, slashed the paintings on the wall, and smashed the mirror in the hallway. Then he took the colour television set. Just as he was about to leave, the tenant returned and surprised him. The intruder struck the tenant and fled with the television set. The tenant was hurt, and he was taken to hospital.

## REFERENCES


Anderson, N.H., and Shanteau, J. Weak inference with linear models. Psychological Bulletin, 1977, 81, 1155-1170.

Akman, D.D., and Normandeau, A. The measurement of crime and delinquency in Canada. British Journal of Criminology, 1967, 6, 129-149.

Akman, D.D., Normandeau, A., and Turner, S. The measurement of delinquency in Canada. Journal of Criminal Law, Criminology and Police Science, 1967, 58, 330-337.

Birnbaum, M.H. The devil rides again: correlation as an index of fit. Psychological Bulletin, 1973, 79, 239-242.

Birnbaum, M.H. Reply to the devil's advocates: don't confound model testing and measurement. Psychological Bulletin, 1974, 81, 854-859.

Comrey, A.L. A proposed method for absolute ratio scaling. Psychometrika, 1950, 15, 317-325.

Corso, J.F. A theoretico-historical review of the threshold concept. Psychological Bulletin, 1963, 60, 356-370.

Ekman, G. Two generalized ratio scaling methods. The Journal of Psychology, 1958, 45, 287-295.

Engen, T., Ross, B.M. Effects of reference number on magnitude estimation. Perception and Psychophysics, 1966, 1, 74-76.

Figlio, R.M. The seriousness of offenses: An evaluation by offenders and non-offenders. Journal of Criminal Law and Criminology, 1975, 66, 189-200.

Good, I.J. Correlation for power functions. Biometrics, 1972, 28, 1127-1129.

Gottfredson, S.D., Young, K.L. and Laufer, W.S. Additivity and interactions in offense seriousness scales. Journal of Research in Crime and Delinquency, 17, 26-41.

Hebart, J.F. Psychologie als wissenschaft, neu gegrundet auf erfahrung, metaphysik, und mathematik. Konigsburg, Germany: Unzer, 1824.

Hsu, M. Cultural and sexual differences on the judgement of criminal offenses: A replication study of the Measurement of Delinquency. Journal of Criminal Law and Criminology, 1973, 64, 248-353.

Kelly, D.H., and Winslow, R.W. Seriousness of delinquent behavior: An alternative perspective. British Journal of Criminology, 1974, 10, 124-135.

Kvalseth, T.O. Seriousness of offenses. Criminology, 1980, 18, 237-244.

Maltz, M.D. Measures of effectiveness for crime reduction programs. Operations Research, 1975, 23, 452-474.

McCleary, R., O'Neil, M.J., Epperlein, T., Jones, C., and Gray, R.H. Effects of legal education and work experience on perceptions of crime seriousness. Social Problems, 1981, 28, 276-289.

McGill, W.J. The slope of the loudness function: a puzzle. In H.R. Moskowitz, B. scharf, and J.C. Stevens (eds.), Sensation and Psychophysics: papers in honor of S.S. Stevens. Boston: D. Reidel Publishing Co., 1974. p. 295-307.

Normandeau, A. The measurement of delinquency in Montreal. Journal of Criminal Law, Criminology and Police Science, 1966, 57, 152-177.

Pease, K., Ireson, J., and Thorpe, J. Modified crime indices for eight countries. Journal of Criminal Law and Criminology, 1957, 66, 209-214.

Pease, K., Ireson, J., and Thorpe, J. Additivity assumptions in the measurement of delinquency. British Journal of Criminology, 1974, 14, 256-263.

Pease, K., Ireson, J., Billingham, S., and Thorpe, J. The development of a scale of offense seriousness. International Journal of Criminology and Penology, 1977, 5, 17-29.

Reidel, M. Perceived circumstances, inferences of intent and judgements of offense seriousness. The Journal of Criminal Law and Criminology, 1975, 66, 201-208.

Robison, S.M. A critical view of the Uniform Crime Reports. Michigan Law Review, 1966, 64, 1031-1054.

Rose, G.N.G. Concerning the measurement of delinquency. British Journal of Criminology, 1966, 6, 414-421.

Rossi, P.H., Waite, E., Bose, C.E., Berk, R.E. The seriousness of crimes: Normative structure and individual differences. American Sociological Review, 1974, 39, 224-237.

Sellin, T., and Wolfgang, M.E. Constructing an index of

delinquency. Philadelphia, Penn., Univ. of Penn., Center of Criminological Research, 1963.

Sellin, T. and Wolfgang, M.E. The measurement of delinquency. New York: John Wiley & Sons, 1964.

Shelley, J.F. Crime seriousness ratings. British Journal of Criminology, 1980, 20, 123-135.

Stevens, S.S. On the Psychophysical Law. Psychological Review, 1957, 64, 153-181.

Stevens, S.S. A metric for social consensus. Science, 1966, 151, 530-541.

Stevens, S.S. and Greenbaum, H.B. Regression effect in psychophysical judgment. Perception and Psychophysics, 1966, 1, 439-446.

Velez-Diaz, A., and Megargee, E.I. An investigation of differences in value judgements between youthful offenders and non-offenders in Puerto Rico. Journal of Criminal Law and Criminology, 1971, 61, 549-553.

Wagner, H., and Pease, K. On adding up scores of offense seriousness. British Journal of Criminology, 1978, 18, 175-178.

Walker, M.A. Measuring the seriousness of crimes. British Journal of Criminology, 1978, 18, 349-364.

Wellford, C.E. and Wiatrowski, M. On the measurement of delinquency. Journal of Criminal Law and Criminology, 1975, 66, 175-188.

Wolfgang, M.E. Uniform crime reports: A critical appraisal. University of Pennsylvania Law Review, 1966, 111, 708-738.

Wong, R. Effects of the modulus on the estimates of magnitude of linear extent. American Journal of Psychology. 1963, 76, 511-512.