Report No. FR-45-U-FAA-86-01

FINAL REPORT

U.S. Department of Transportation
Research and Special Programs Administration
Transportation Systems Center
Kendall Square
Cambridge, MA 02142

ANALYSIS AND TESTING OF KOORNSTRA-TYPE INDUCED EXPOSURE MODELS

Peter Mengert

October 1985

SPONSOR: Ernst Meyer, NRD-31
U.S. Department of Transportation
National Highway Traffic Safety Administration
Mathematical Analysis Division

APPROVED FOR DISTRIBUTION: _____
Chief, Operator Performance
and Safety Analysis Division

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Induced exposure models postulate a structure for accident data which permits the estimation of two factors: exposure and proneness. Since information on exposure is needed in order to assess the accident risk of different driver, vehicle, and environmental situations and since reliable exposure data is expensive to collect, induced exposure models hold the promise of a rich exposure data source which is perfectly matched to the accident data to be analyzed. This paper assesses the validity of the postulated structure of one induced exposure model: the Koornstra Model. The Koornstra Model was chosen for analysis and testing because it appeared to have the most potential for usefulness based on :

a) previously reported favorable results (in limited testing);
b) universal applicability;
c) damaging criticism of certain other models; and
d) a rich and well posed model structure.

This paper analyses and tests the Koornstra Model from three different points of view:

1) Is it based on reasonable assumptions?
2) Does the model provide a significantly better fit to accident data than a simpler model which does not permit exposure or proneness to be estimated?
3) How do the exposure estimates provided by the model compare with those from externally collected data?

Assumptions:

When the assumptions needed to derive the model are expressed in words, the key assumptions can be reduced to the following:

1) All driver groups have the same proportion of their driving in any time and location. Further, encounters with other drivers are distributed the same for each driver group.
2) Accident fault is primarily a single driver phenomenon and does not involve unsafe moves by both drivers.

3) Relative driving safety of a driver (proneness) is largely constant across accident situations.

Each of these assumptions is violated in the real world, but it may be possible to structure the accident data so that they hold adequately. For example, if the first assumption is violated the data can be disaggregated by time and location in an attempt to improve the validity of the assumption.

External Exposure:

The Basic Koornstra Model exposure estimates were compared with externally estimated exposure estimates for Ulster County, New York, which was of questionable accuracy. The agreement was found unsatisfactory in each case. When accident rates derived from the induced exposure estimates were compared with accident rates derived from direct exposure estimates the agreement was also very poor in every case.

Significant Fit:

The fit of the Basic Koornstra Model was compared to the fit of a simpler model in explaining accident data from Ulster County, New York and North Carolina. The simple model does not allow exposure to accident situations to be separated from proneness. Though many stratifications of each data set were tested, no evidence was found to suggest that the Koornstra model fit better than the simpler model, leading us to conclude that the pronesness and exposure estimates from the Basic Koornstra Model cannot be determined with any acceptable statistical stability. Or, viewed somewhat differently, there is no internal evidence that the more complex Basic Koornstra Model fits better than the simple multiplicative model, and therefore, the Basic Koornstra Model is rejected as unsuitable for the data. In the testing on the limited accident sample for Ulster County, New York, it could be argued that the Basic Koornstra Model did not fit significantly better because the accident sample of about 900 two-car accidents was too small. However, tests on the much larger North Carolina accident data of over 100,000 two-car accidents yielded the same results even when the data was extensively disaggregated by time and location.

This study also included more limited testing of the One-and-Two-Car Koornstra Model and the related Thorpe Model, both of which had been criticized in earlier papers. These were rejected in this study as not giving credible exposure estimates

when applied to the Ulster County data. It has been noted in the literature that the Thorpe Model overestimates the exposure of older drivers and underestimates that of younger drivers. Because of its close connection to the Thorpe Model, the One-and-Two-Car Koornstra Model does the same thing.

These results all show that the Koornstra type induced exposure models are of little practical value in producing useful exposure estimates from highway accident data.

1.0 INTRODUCTION

The need for data on the exposure of driver and vehicle groups to highway accident producing situations has long been recognized. Accident statistics which cite accident rates per mile or per registered vehicle are attempts to control for the different exposure to accidents of groups of drivers and vehicles. Most sources of exposure data are limited in the extent to which they identify groups of drivers, vehicle and situations. As safety analysts continue to press for finer distinctions there is a temptation to assume that certain types of accidents are more the result of presence in the traffic stream, i.e., of exposure while other accidents are more influenced by a particular quantity to be called "proneness" in this report. The meaning of proneness and of exposure will be developed in more detail especially in Chapter 2.0. That this conceptual step can probably lead to useful estimates in certain cases may lead one to hope that systematic procedures could be developed to obtain accurate exposure estimates from accident data in a wide variety of circumstances. Any procedure for getting exposure estimates from accident data may be referred to as an induced exposure model. Several rather general induced exposure models and procedures have been suggested in the last two decades and have been tested to varying extents.

This report focuses on the Koornstra Model because it appears to have the most potential for application and has clear modeling assumptions. A description of how a Koornstra Model can extract accident exposure and accident proneness information from accident data is presented below. It is followed by a brief summary description of other induced exposure models and finally by a statement of the objectives of this study.

1.1 INFORMAL JUSTIFICATION FOR THE USE OF THE ACCIDENT MATRIX TO DETERMINE EXPOSURE AND PRONENESS

The Koornstra Model attempts to estimate exposure and proneness values pertaining to classes of drivers by analyzing the accident involvement matrix X_{ij} containing counts of observed collision involvements for drivers of group i and

group j* (the driver groups would ordinarily be defined by some classification such as by age and sex). A rough idea of the principle behind such an analysis can be obtained by the following considerations. Suppose that in addition to several groups of ordinary drivers there are two special groups of drivers, one group called the "super good drivers" and the other called the "super bad drivers." The super good group will be labelled group G, and the super bad group B, and the ordinary drivers will be groups 1 and 2.

The super good drivers are so good that when an ordinary driver (from group 1 or 2) has a collision with a member of group G the accident is almost certainly caused by the ordinary driver. On the other hand, when an ordinary driver has a collision with a member of group B, the accident is almost surely caused by the group B driver (super bad).

The collisions which ordinary drivers have with members of group B are mostly due to being at the wrong place at the wrong time and are thus largely a measure of exposure. The collisions with members of group G are due to unlucky circumstances in part but for the most part can happen only if there is fault on the part of the driver (from the ordinary group). Thus, these accidents are proportional to both exposure and proneness.

Table 1 gives a hypothetical accident involvement matrix for groups G, 1, 2, and B. From this data the ratio of exposures of groups 1 and 2 can be estimated as $X_{1B}/X_{2B} = 2.22$ since X_{1B} is the number of collisions of group 1 with super bad drivers and thus is assumed to be proportional to the exposure of group 1 and similarly X_{2B} proportional to the exposure of group 2. Their ratio is thus an estimate of the ratio of exposures and shows that group 1 has a little over twice the exposure of group 2.

The ratio of exposure times proneness is estimated by

$$X_{1G}/X_{2G} = .53$$

*Thus, for example, X_{12} would be the number of collisions in which one driver belonged to group 1 and the other to group 2.

TABLE 1
EXPOSURE FROM ACCIDENT DATA: AN EXAMPLE

X_{ij}

	G	1	2	B
G	60	160	300	600
1		100	120	200
2			80	90
B				40

X_{B1}/X_{B2} = ratio of exposure = 200/90 = 2.22

X_{G1}/X_{G2} = ratio of exposure times proneness = 160/300 = .53

ratio of proneness = .53/2.22 = .24

Thus the ratio of proneness (group 1 to group 2) is estimated by .53/2.22 and this shows that group 2 has about 4 times the proneness of group 1.

Given the existence of super good and super bad drivers (and a knowledge of which is which) the relative exposure and proneness of each group of ordinary drivers is easily obtained from the number of collisions with the super good and super bad drivers. However, if there are no super good and super bad drivers, similar information can be obtained from the accident matrix of ordinary drivers. Each accident experienced by a group of drivers increases both the estimated exposure and the estimated product of exposure and proneness (for shortness of expression this product will be referred to as the accident potential). If the collision is with a group of relatively bad drivers, there will be more of an increase in the estimated exposure. If, on the other hand, the collision is with a driver from a group of relatively good drivers, the estimated accident potential will increase more than the estimated exposure.

Consequently, the complete accident matrix for a goup of ordinary drivers gives information on their relative exposure and proneness values. There appears to be some circularity in these considerations because one must know which groups are good drivers and which are bad drivers before one can compute the proneness values for other groups. The Basic Koornstra Model admits two symmetric solutions in which the exposure and accident potential are interchanged. In a given situation at most one of the solutions will make any sense if some prior knowledge of relative proneness and exposure is available. This prior knowledge could be as little as a rough ranking on the basis of exposure or of proneness. Thus the ambiguity is easily resolved in most practical situations and the circularity is broken as well.

1.2 OTHER INDUCED EXPOSURE MODELS

In addition to the Koornstra Model which is described in detail in Chapter 2.0, two other types of models have been suggested. The Thorpe* model derives the relative exposure and proneness of a driver group from its proportion, s, of the single car accidents and its proportion, t, of the two car (collision) involvements.

*See Reference 2.

The Thorpe model estimates the relative exposure, e_i, of the driver group, i, as: $e_i = 2\ t_i - s_i$, the theory being that two car accidents are only partially the fault of the given driver while single car accidents are completeley the fault of the driver group involved. Several assumptions were needed to derive the model, but one assumption has been found to be severely violated in the case of driver age groups and is the key assumption of Thorpe's model considered in this paper. This is the assumption that relative proneness is the same for single and two car accidents. The evidence which contradicts this assumption is that older drivers (over 50 years old) appear to have relatively lower proneness to single car accidents than to two car accidents.

The Haight models are based on single car vs. two car accidents as is the Thorpe Model. They were rejected from further consideration on this project on the basis of this and other considerations as discussed in Reference 1.

Assigned responsibilitly models have been considered by Carr, Hall, and Cerelli among others.* They have been referred to as "quasi-induced exposure" models by Haight.** These models use assigned responsibility (by police citation or by accident circumstances) to separate proneness from exposure. Specifically, relative exposure is measured by relative involvement as the not-responsible party in two car collisions.

These models have been investigated several times with consistently positive results. However, there is relatively little evidence derived from direct testing of the models. In general, the hypothesis that carefully chosen assigned responsibility information works well for computing exposure from accident data appears tenable.

It is assigned responsibility data which refutes Thorpe's hypothesis and shows that the Thorpe model can give very inaccurate and misleading estimates of exposure. Carr gave dramatic evidence of this and the conclusion was confirmed by other workers in quite different contexts.

*References 6, 8, and 9.

**Reference 7.

1.3 OBJECTIVE

The objective of this study is to assess the usefulness of Koornstra induced exposure models in accident analysis. There are three components to this assessment. First, the basic assumptions needed to derive the Koornstra Model are identified and alternatives are suggested. The reasonableness of the assumptions and the alternatives represents one level of assessment. It is presented in Chapter 2.0. Second, methods of testing and evaluating Koornstra Models on accident data are developed in Chapter 3.0. Third, the Koornstra Model is tested on accident data from Ulster County, New York and on data from North Carolina. The results of these tests are presented in Chapter 4.0.

2.0 EXAMINATION OF KOORNSTRA MODELS

In this chapter, the Basic Koornstra Model is described (2.1); other, more general Koornstra Models are discussed (2.2); the assumptions leading to the Koornstra Model are specified (2.3); and the Koornstra Model is derived (2.4).

Koornstra proposed a model and it was applied to traffic accident data from the Netherlands. The Koornstra model also was the main subject of a rather large investigation on Danish data reported by Wass (Reference 4). Wass investigated only the One-and Two-Car Koornstra Model* even though Koornstra had concluded (in agreement with the discussion below in this report) that the Two-Car Koornstra Model (or Basic Koornstra Model) was more promising. Wass concluded that the Koornstra model worked quite well on the Danish data. This conclusion will be seen to be at odds with the conclusions of this report.

2.1 THE BASIC KOORNSTRA MODEL

In introducing the Koornstra type models, it is best to start with what can be called the Basic Koornstra Model. Koornstra's original paper considered a more complex model, but the simpler model introduces the fundamental concepts. It also introduces fewer assumptions which may not be valid. Later an extended model (the One- and Two-Car Model) given by Koornstra will also be introduced, as will another extension (the Generalized Koonstra Model) useful in analyzing the applicability of the Basic Koornstra Model.

The Basic Koornstra Model takes as data the accident involvement matrix X_{ij}. It is helpful at the outset to also define the accident matrix A_{ij} for comparison, but all subsequent discussions will refer to X_{ij}. The accident matrix A_{ij} is the number of collisions between members of group i and group j. In the initial discussion of the model it is assumed that the groups are classes of drivers (the case where they are classes of vehicles or driver-vehicle combinations may also be considered). A_{12} is the number of observed collisions in which one driver was in group 1 (say a particular age-sex category) and the other driver was in

*See next section for definition of various Koornstra type models.

group 2. The involvement matrix X_{ij} is defined as the number of involvements that drivers in group i have in collisions in which the other driver was in group j. Then $X_{ij} = A_{ij}$ if $i \neq j$ but $X_{ii} = 2A_{ii}$. In other words, involvements equal accidents when the other party is different but since each accident produces two involvements there are twice as many involvements within a group as there are accidents. Since each accident leads to two involvements:

$$\sum_k \sum_j X_{kj} = 2N$$

where N is the total number of two car collisions. (In terms of A_{ij} we have

$$\sum_k \sum_{j \leq k} A_{kj} = N.$$

The Basic Koornstra Model states:*

$$X_{ij} \approx \hat{X}_{ij} = (p_i + p_j)\, e_i e_j$$

The quantity e_k is taken to measure relative exposure of class k and p_k measures its relative proneness. These quantities will be discussed in more detail below.

The Basic Koornstra Model is used to estimate the p_k's and e_k's given the X_{ik}'s. First, one makes the fairly standard assumption that A_{ij} is Poisson distributed with mean $\hat{A}_{ij}$ ($\hat{A}_{ij} = \hat{X}_{ij}$ if $i \neq j$, $\hat{A}_{ii} = 1/2\ \hat{X}_{ii}$) and then finds maximum likelihood estimates for the e_k's and p_k's. Some other goodness of fit criterion could be maximized with respect to the model parameters (e_k's and p_k's) but maximum likelihood is probably the best. (Koornstra suggested minimizing a chi square statistic which is nearly equivalent to maximizing the likelihood estimate for large samples). More is said on the statistical accuracy of this process in Reference 5 which also describes the details of finding the maximum likelihood solution.

A question which naturally arises is why should the e_k (as estimated by this procedure) be expected to be good measures of exposure and p_k of proneness (proneness in the sense of probability of having an accident given exposure to the accident situation). More fundamentally why does the involvement matrix give any information at all on exposure and proneness separately.

*Koornstra's original paper used the accident matrix rather than the involvement matrix here. The involvment matrix must be used, however. As just seen the involvement matrix is easily calculated from the accident matrix.

These two questions are addressed in Sections 1.1 and 2.3. Section 1.1 informally dealt with the second question regarding how evidence about the amount and safety of driving can be present in the accident matrix. Section 2.3 gives a derivation of the Generalized Koornstra Model defined in Section 2.2, carefully identifying each assumption as invoked in the derivation. It is the reasonableness of these assumptions which lead to confidence that e_k is exposure and p_k is proneness.

Then, the assumptions needed in deriving this model and the more restricted Basic Koornstra Model are discussed and their implications concerning validity and testing are considered in subsequent sections. It should be noted that even if

$$X_{ij} = \hat{X}_{ij}$$

is a good approximation as asserted by the Koornstra model, the questions whether p_k and e_k estimate exposure and proneness still remain. However, if all the assumptions in the deriviation hold then e_k and p_k must measure exposure and proneness. These matters will be dealt with more fully in subsequent sections.

2.2 OTHER KOORNSTRA-TYPE MODELS

Besides the Basic Koornstra Model two other related models will be referred to in this report. The first is the One- and Two-Car Koornstra model. This model was introduced in Koornstra's original paper and was the model of primary interest in that paper. It was also the model of primary concern for Wass. This model has the form of the Basic Koornstra Model but also includes (at least) one category of fictitious "drivers" to represent accidents which don't involve collisions with other vehicles. X_{io} is the number of single car accidents involving driver class i. By assumption $\hat{X}_{oo} = 0$ and $\hat{p}_o = 0$. In the maximum likelihood solution, the condition $X_{oo} = 0$ ensures that the estimate of $\hat{p}_o$ is exactly zero.

The One- and Two-Car Koornstra Model is, in effect, a combination of the Basic Koornstra Model and the Thorpe Model (the formal aspects of this assertion are discussed in Reference 1). In practice the estimates are strongly affected by the

comparison of One- and Two-Car accident counts. This is the crux of the Thorpe Model. It has been found that the assumption of equal proneness in One- and Two-Car accidents is badly violated in the case of driver age groups and so the Thorpe Model is probably seldom of any value in estimating proneness and exposure. It is therefore suggested that the One- and Two-Car Koornstra Model is less likely to be valid than the Basic Koornstra Model. This is one of the main conclusions of Koornstra's second paper.

The second extension of the Basic Koornstra Model is the Generalized Koornstra Model.

The Generalized Koornstra Model can be expressed:

$$\hat{X}_{ij} = (p_i + p_j + \alpha p_i p_j + \beta) e_i e_j \quad (\alpha \geq 0, \beta \geq 0)$$

Simple algebraic manipulation shows that X_{ij} as given by this expression can also be expressed thus:

$$\hat{X}_{ij} = (p'_i + p'_j) e'_i e'_j$$

providing $\alpha/\beta \leq 1$.

Of course p'_i and e'_i are mixed functions of p_i and e_i (p'_i does not depend on p_j or e_j ($j \neq i$), nor does e'_i). Consequently, p'_i is not a direct measure of proneness if p_i is, and e'_i is not a direct measure of exposure if e_i is such a measure (in Section 3.5 it will be shown that the Generalized Koornstra Model results from certain more general assumptions than used in the Basic Koornstra Model). However, the p'_i's will be in the same rank order as the p_i's. The variation in the p'_i's will be less than that in the p_i's (see Reference 1 for fuller discussion of these points).

Consequently, p'_i and e'_i will have the fortunate property that they assign the accident variation more to exposure than to proneness variations. Thus, they lead to conservative estimates of differences in accident rates (providing of course that e_i is true exposure).

The Generalized Koornstra Model can not be applied to data to estimate p_i and e_i. It is introduced for the purpose of deriving the Basic Koornstra Model, and even more importantly, for providing the basis for analyzing the results of

applying the Basic Koornstra Model. The key point to keep in mind is that if $\alpha_{kj} = 1$ it is not possible to separate proneness from exposure even if the data are governed by a Generalized Koornstra Model.

2.3 ASSUMPTIONS NEEDED TO DERIVE THE KOORNSTRA MODEL

The aim of this section is to present two sets of assumptions sufficient to derive the Basic Koornstra Model, show that the model can be derived from the assumptions, and discuss the significance of the assumptions and what circumstances can affect their validity.*

The derivation and the assumptions needed focus on what will be called "accident situations" (or "situations" for short), each involving two drivers. The specificaton of a situation requires a specification of a point in space and time and a scenario. Thus, a situation encompasses a particular time on a particular day, a particular spot on a particular road and a particular traffic pattern (especially as regards the two vehicles driven by the two drivers to which the situation refers). It is in the nature of these situations that no accident ever occurs except in conjuction with one of these (conceptual) situations.** The situation does not include a specification of the two drivers (or vehicles) involved.

The expected number of accidents between two specific drivers (say driver k and driver j) is equal to the sum of the expected number in each situation. Since a situation is a brief occurrence, only one accident can occur and the expected number of accidents in a given situation is equal to the probability of an accident. The result is that the expected number of accidents (between drivers k and j) is the sum over all situations of the probabilities of an accident between k and j in that situation.

*The assumptions and derivations given here are quite different from those in the original Koornstra paper and in the very similar treatment by Wass. However, the assumptions given here are sufficient to derive the model and are substantially easier to interpret than Koornstra's assumptions.

**In this sense, the concept involved is somewhat similar to that of "conflict" which is sometimes invoked in accident analysis.

The probability that a collision will occur between drivers k and j in the situation s will be denoted by $Pr(C_{kjs} | k,j,s)$. Here C_{kjs} symbolically represents a collision between k and j in situation s. The accident can occur only if drivers k and j are present at situation s. This event will be donoted by E_{kjs}. Then by the rules of conditional probability

$$Pr(C_{kjs} | k,j,s) = Pr(C_{kjs} | E_{kjs}, k,j,s) \; Pr(E_{kjs} | k,j,s)$$

The Koornstra model is derived by making simplifying assumptions about these factors: The first assumption deals with the factor $Pr(E_{kjs} k,j,s,)$:

Assumption 1:

$$Pr(E_{kjs} | k,j,s,) = r_s \, e_k \, e_j$$

This assumption can also be stated*

$$Pr(E_{kjs} | k,j,s) = g_{js} \, e_k$$

The previous expression follows immediately from the fact that the same form for the expression must hold when k and j are interchanged.

In words, the practical content of Assumption 1 can be expressed as follows:

> "The exposure of each driver is distributed the same as that of all other drivers over space and time and encounters with other drivers."

Equal distribution of exposure over space and time simply means that the probability of a given driver's being present in a given situation is proportional only to his overall exposure (proportional to e_k for driver k), and that the proportion is the same for all drivers! Violation of this assumption is called incomplete mixing** because some drivers must have more of their driving in a situation than other drivers. Equal distribution of drivers over encounters with other drivers would follow from equal distribution over space and time were it not for the effect of different speeds on the highways. In Appendix E the effects of differential speeds on numbers of encounters between vehicles is discussed. This is a very complicated matter and will not be discussed in detail in the text.

*Note that r_s and g_{js} are arbitrary functions of their subscripts. The assertion concerns how the probability depends on s, j, and k.

**A term introduced by Haight in Reference 7.

The next assumption deals with the other factor, $Pr(C_{kjs} | E_{kjs}, k,j,s)$. There are two possibilities for the second assumption each of which is sufficient to derive the Koornstra model (when taken in conjunction with Assumption 1). Both will be considered because neither can be derived from the other and each might be expected to be approximately true in certain circumstances:

Assumption 2a:

$Pr(C_{kjs} | E_{kjs}, k,j,s) = p_{ks} + p_{js}$

In words:

"The probability of an accident given that drivers k and j are present in situation s is equal to the sum of two terms each depending on the situation and one driver only."

What is seriously missing here is a product term of the form $q_{ks}\, q_{js}$. The product term would arise if one calculated the probability of an accident occurring due to actions on the part of both drivers.

The alternate assumptions is not missing the cross product term (it however, requires other restrictive conditions):

Assumption 2b1:

$Pr(C_{kjs} | E_{kjs}, k,j,s) = (a_s p_k + b_s p_j + c_s p_j p_k + d_s)$

Assumption 2b2:

For each s there is an s' such that*

$Pr(C_{kjs}) = Pr(C_{jks'})$

Assumption 2b1 and 2b2 together with assumption 1 allow the Generalized Koornstra model to be derived.

A further assumption now allows the Basic Koornstra Model to be derived from the Generalized Koornstra model.

The derivation of the Basic Koornstra Model from assumptions 1 and 2a and the derivation of the Generalized Koornstra model from assumptions 1 and 2b are

*For every situation the reverse situation occurs with equal frequency.

quite simple and given in the subsection following this one. No further assumptions are required; the procedure is strictly mathematical.

Assumption 2a can be stated in words approximately as follows;

> "Accidents are essentially the result of fault by one party or the other, or neither, but not the fault of both."

This is meant to say that an unsafe action is required at most on the part of one driver.

This assumption is not absolutely required since assumption 2b can be substituted. It says in effect "There is a proneness for each driver, constant over situations, such that the accident probability is always linear in this proneness." A weaker statement that probably contains the essential requirements is: "If one of two drivers is safer (has lower proneness) in one situation, then that driver is safer in any other situation." Since 2b yields only the Generalized Koornstra Model another assumption is needed with 2b to yield the Basic Koornstra Model. It was explained in Reference 1 that the Generalized Koornstra Model

$$X_{kj} \cong \hat{X}_{kj} = (p_i + p_j + \alpha p_i p_j + \beta)\, e_i e_j$$

can be transformed into a Basic Koornstra Model

$$X_{ij} = (p'_i + p'_j)\, e'_j e'_j$$

if $\alpha\beta \leq 1$. The proneness and exposure values are mixed; however, the proneness values do not change their rankings and the estimated proneness values show less variation than the true proneness values. In this sense applying the Basic Koornstra Model leads to conservative estimates. The key extra assumption is:

> 2b3: $\alpha\beta < 1$

In words this could be expressed roughly:

> "Two-Car accidents are usually and to a large degree caused by overriding fault of one driver rather than of both drivers (or neither)."

This is clearly very similar to assumption 2a but is weaker.

In summary the key assumptions involved in 2a and 2b are (in words)

i) "Accident fault is primarily a single driver phenomenom and to a large extent does not require unsafe moves on the part of both drivers."

ii) "Relative driving safety (as measured by proneness) is largely constant across two-car accident situations."

Clearly all the assumptions mentioned in this section can be violated to some degree. The key question is not whether they are absolutely valid but whether they hold to the degree needed to make the Basic Koornstra Model useful for estimating exposure from accident involvement matrices.

It is suggested that Assumption 1 is the most critical for the validity of the Koornstra model. Some more will be said about this in Section 3.6.

This section is ended with examples of how the assumptions are violated:

Assumption 1: This assumption is violated when certain driver groups get a larger percentage of their exposure at certain times (e.g., nighttime) than other groups or when certain groups of drivers do more of their driving on certain types of roads (e.g., rural roads) than other groups of drivers.

The degree to which this assumption holds can always be improved by stratifying the data (for example to only daytime accidents).

Differences in speeds between driver groups can also violate this assumption as mentioned earlier and discussed in Appendix E in more detail.

Assumption 2a: This assumption is violated for accidents which occur only due to mistakes on the parts of both drivers e.g., one driver goes through a stop sign without stopping and another driver collides with him due to inattentiveness (i.e., avoidably).

Assumption 2b: There is very good evidence that older drivers are relatively safe drivers in certain single car accident situations but relatively unsafe in certain two car accident situations. This would suggest that a reversal in proneness rankings is possible within two car accident context. For example an older driver may be less likely than a younger driver to stray into the oncoming traffic lane but may be less skillful in avoiding another driver who strays into his lane.

2.4 DERIVATION OF THE KOORNSTRA AND GENERALIZED KOORNSTRA MODELS FROM GIVEN ASSUMPTIONS

The first derivation to be given in this section is of the Basic Koornstra Model from assumptions 1 and 2a (see Section 3.4 for notation and preliminary discussion). The expected number of accidents between drivers k and j is

$$\hat{X}_{kj} = \sum_s \Pr(C_{kjs} \mid k,j,s) = \sum_s \Pr(C_{kjs} \mid E_{kjs}, k,j,s)\ \Pr(E_{kjs} \mid k,j,s)$$

By assumption 1 an 2a this is

$$\hat{X}_{kj} = \sum_s (p_{ks} + p_{js})\, r_s\, e_k\, e_j$$

$$= (p_k + p_j)\, e_k\, e_j$$

where

$$p_k = \sum_s r_s\, p_{ks}$$

Applying the aggregation Theorem (see Appendix A) the result is

$$\hat{X}_{ij} = (p_i + p_j) e_i e_j$$

where now i and j refer classes of drivers instead of individual drivers,

$$e_i = \sum_{k \in i} e_k \text{ and } p_i = \sum_{k \in i} p_k\, e_k \Big/ \sum_{k \in i} e_k$$

i.e., e_i is the sum of the exposures of the individual drivers in class i and p_i is the weighted average of proneness values for drivers in class i (with the weighting based on exposure).

In order to apply the aggregation theorem to $\hat{X}_{kj}$ it is necessary to assume that

$$\hat{X}_{kk} = (p_k + p_k) e_k e_k = 2 p_k e_k$$

while in reality it is impossible for an individual driver to collide with himself

and therefore $X_{kk} = 0$. Since, however, there are so many drivers in the total population this introduces a negligible error into the estimate and so the derivation of the Basic Koornstra Model from assumptions 1 and 2a is complete.

In a very similar manner the Generalized Koornstra Model can be derived from assumption 1 and assumption 2b. In abbreviated form the derivation is as follows:

$$\hat{X}_{kj} = \sum_s \Pr(C_{kjs} \mid k,j,s)$$

$$= \sum_s \Pr(C_{kjs} \mid E_{kjs}, k,j,s)\ \Pr(E_{kjs} \mid k,j,s)$$

$$= \sum_s (a_s p_k + a_s p_j + c_s p_k p_j + d_s)\ r_s e_j e_k$$

(since by 2b2, $b_s = a_s$)

$$= C(p_k + p_j + \alpha\, p_k p_j + \beta)\ e_k\ e_j$$

where $C = \sum_s a_s r_s$

$$\alpha = \sum_s c_s r_s$$

$$\beta = \sum_s d_s r_s$$

The factor C can be incorporated into the e_k's by letting $e_k = e_k\sqrt{C}$, $e_j = e_j\sqrt{C}$ so that

$$X_{kj} = (p_k + p_j + \alpha\, p_k p_j + \beta)\ e_k\ e_j$$

Now the aggregation theorem yields

$$\hat{X}_{ij} = (p_i + p_j + \alpha\, p_i p_j + \beta)\ e_i\ e_j$$

where

$$e_i = \sum_{k \in i} e_k$$

and

$$p_i = \sum_{k \in i} p_k\, e_k \Big/ \sum_{k \in i} e_k$$

just as the previous case where the Basic Koornstra Model was derived for assumptions 1 and 2a. This completes the derivation of the Generalized Koornstra Model from Assumption 1 and 2b.

3.0 AN APPROACH TO TESTING AND APPLYING THE KOORNSTRA MODEL

This section describes a two-step technique to be used in applying and testing the Koornstra model. The focus of this section is on techniques for judging the adequacy of the model which can be applied to the accident data alone, since this is the situation in which most induced exposure models would be applied. The techniques first examine the structure of the accident involvement matrix using eigen value/eigen vector* analysis, and second examine the adequacy of the fit in relation to a simpler model which does not permit exposure and proneness to be separated. Unless the Koornstra Model fits better than this simple model, it cannot provide information on exposure and proneness.

The first technique to be applied is an eigen value/eigen vector analysis of the accident involvement matrix. This is an exploratory technique to determine if the data suggest that the Basic Koornstra Model model fits.

To see how the eigen value/eigen vector analysis applies, first note that the Generalized Koornstra Model,

$$X_{ij} = (p_i + p_j + \alpha p_i p_j + \beta)\, e_i\, e_j$$

is equivalent to

$$X_{ij} = \lambda_1 U_i U_j + \lambda_2 V_i V_j$$

where $\underline{U}$, $\underline{V}$ are orthonormal vectors (i.e., $\sum_i U_i V_i = 0$, $\sum_i U_i^2 = 1$, $\sum V_i^2 = 1$)

Thus λ_1 and λ_2 are the only (non-zero) eigen values of X_{ij} and $\underline{U}$ and $\underline{V}$ are its eigen vectors. By choice $|\lambda_1| > |\lambda_2|$. Since X_{ij} is positive $\lambda_1 > 0$. The critical point is whether $\lambda_2 > 0$ or $\lambda_2 < 0$.

If $\lambda_2 < 0$ then the model is equivalent to the Basic Koornstra Model (this corresponds to $\alpha\beta < 1$ in the Generalized Koornstra Model).

If $\lambda_2 = 0$ a simple multiplicative model holds e.g., $X_{ij} = e_i\, e_j$ or $X_{ij} = e_i p_i\, e_j p_j$. In this case no separate proneness and exposure quantities can be calculated for each driver class.

*There are many references for this--especially books on applied linear algebra or numerical analysis, e.g., Reference 10.

If $\lambda_2 > 0$ then the Generalized Koornstra Model does not reduce to the Basic Koornstra Model. This is equivalent to $\alpha/\beta > 1$ in the Generalized Koornstra Model. It may suggest that Assumption 2 (of Sec. 2.4) does not hold and that accidents in which both parties contribute fault are important. It may instead (or in addition) suggest that the mechanism which produces a non-multiplicative form for X_{ij} is incomplete mixing (violation of Assumption 1) rather than variable proneness.

To see how incomplete mixing leads to a positive second eigen value consider a "mixed multiplicative" model. In a mixed multiplicative model all drivers have the same proneness but their exposures are distributed differently over space and time (in violation of assumption 1). For example let e_{1k}, e_{2k}, e_{3k} be the exposure of driver k in time periods 1, 2, 3. Then if all proneness are equal (say each equals 1), the model

$$X_{ij} = \sum_r e_{ri}\, e_{rj}$$

results. It is straightforward to show that X_{ij} in this case has only positive eigen values. It may also have more than two non-zero eigen values.

In summary, if X_{ij} has a positive second eigen value, a strong breakdown of assumption 1 or assumption 2 or both is suggested. No application of the Koornstra Model is possible, in this case.

The exploratory eigen value/eigen vector analysis proceeds as folows: Form the eigen values of the matrix X_{ij}. The largest eigen value (in absolute value), λ_1, will be positive. The second largest eigen value is λ_2.

1. If $\lambda_2 \geq 0$ then the Koornstra model is not appropriate to this data.
2. If $\lambda_2 < 0$ then the Koornstra model may be applicable. The key question is whether λ_2 is significantly less than zero. That question is addressed at the next stage.

Finally (for the Koornstra model to hold) the other eigen values of X_{ij} (3 , 4 , . . .) should all be considerably less in absolute value than λ_2. If λ_3 is comparable in magnitude to λ_2 and of opposite sign, there is evidence that the sign of λ_2 is not statistically significant. λ_2's difference from zero is considered next. Thus,

the eigen value analysis of X_{ij} is primarily to determine the sign of λ_2. If λ_2 is positive the procedure is ended because there is no prospect of applying the Koornstra model. If λ_2 is negative, the analysis proceeds to the next step.

In this step the fit of the Basic Koornstra Model is compared to the fit of the Simple Multiplicative Model*. The Simple Multiplicative Model is $X_{ij} = p_j e_j p_i e_i$. In the Simple Multiplicative Model exposure cannot be separated from proneness. So, if the Koornstra model does not fit significantly better then there is inadequate information in the involvement matrix to estimate both proneness and exposure. The Basic Koornstra Model

$$X_{ij} = (p_i + p_j)\, e_i e_j$$

is fit to the data (accident involvement matrix X_{ij}) using a maximum likelihood procedure. If a Poisson distribution for accident counts is assumed, then the log likelihood was shown in Reference 5 to be:

$$L = 1/2 \sum_{ij} X_{ij} \log(\hat{X}_{ij}/X_{ij}) - \hat{X}_{ij} + X_{ij}$$

excluding terms which do not involve $\hat{X}_{ij}$ (such terms may be ignored when maximizing over $\hat{X}_{ij}$). (This expression for L is chosen since L = 0 if $\hat{X}_{ij} = X_{ij}$, otherwise L is negative). In Reference 5 it was also indicated how sampling variances of the estimated parameters e_k and p_k could be determined.

The next step after fitting the Basic Koornstra Model to the data is to compare the quality of fit to that of the Simple Multiplicative Model:

$$X_{ij} \simeq \hat{X}_{ij} = W_i W_j$$

where a maximum likelihood (Poisson) estimate of W_i can be chosen simply as

$$W_i = \sum_j X_{ij} / \sum_{jk} X_{jk}$$

The X^2 value:

$$1/2 \sum_{ij} (X_{ij} - \hat{X}_{ij})^2 / \hat{X}_{ij}$$

is computed for both the Basic Koornstra Model and the Simple Multiplicative Model model.

The number of degrees of freedom (independent parameters) in the Basic Koornstra Model for N classes of drivers is 2N - 1 (the -1 term is there because an arbitrary constant can multiply all the p_k's and divide all the e_k's). The number of independent parameters in the Simple Multiplicative Model is N. The

*A similar approach to testing the applicability of the Basic Koornstra model appears in Reference 11.

number of independent data cells is $(1 + 2 \ldots + N) = N(N + 1)/2$. The residual degrees of freedom when fitting the Koornstra model is thus $N(N+1)/2 - (2\ N-1)$. When fitting the Simple Multiplicative Model it is $N(N+1)/2 - N$. The difference is N-1. The amount by which the X^2 statistic is smaller for the Basic Koornstra Model than for the Simple Multiplicative Model is compared to the difference in degrees of freedom, (N-1). The difference in X^2 is referred to a standard X^2 (chi square) table with degrees of freedom equal to N-1. If the difference in X^2 is not significantly large for the given difference in degrees of freedom, it is concluded that the Basic Koornstra Model does not fit significantly better than the Simple Multiplicative Model. If the X^2 difference is significant then the Basic Koornstra Model does fit significantly better than the Simple Multiplicative Model.

In this case the chi square value for the Basic Koornstra Model should also be referred to its residual degrees of freedom $N(N+1)/2 - (2N-1)$ to see if the overall fit is satisfactory. If the overall chi square is not significantly large for the degrees of freedom, the fit is good. However, if the chi square value is singificantly large, there is evidence that the fit is poor.

Regarding the process of fitting the Basic Koornstra Model to the data by maximizing the Poisson likelihood, two further points should be raised here:

> The first concerns the situation in which the second eigen value is negative but the Basic Koornstra Model does not fit significantly better than the Simple Multiplicative Model. This might be due to an insufficiently large sample size in conjunction with proneness values which do not vary much between driver classes. Another and perhaps more likely explanation is that incomplete mixing cancels out proneness. Specifically, the following combination of conditions can arise: Even though the Basic Koornstra Model holds on separate strata (of time and/or space) and proneness varies substantially among driver groups, the different exposure distributions over the strata result in an accident involvement matrix best fit by a simple multiplicative model. In this case, proneness and exposure values cannot be disentangled by an analysis of the accident involvement matrix and a

larger sample size will not help (except if it permits more appropriate strata of accident situations).

The second point concerns the case where the second eigen value of the accident involvment matrix is positive. In this case, what happens when a maximum likelihood fit of the Basic Koornstra Model is obtained, which entails a negative second eigen value? As might be expected, the model degenerates into the Simple Multiplicative Model, with its zero second eigen value.

4.0 EMPIRICAL TESTING

Because of an ongoing study in Ulster County, New York which involved the direct collection of exposure (quantity of driving) data categorized by age, sex and other variables and because accident data were also available for Ulster County, it was decided to test the Koornstra and Thorpe models on the Ulster County accident data. This test is described in Section 4.1 below.

The data to be used consists of all accidents in the New York State accident file pertaining to Ulster County during the study period.

Since the Ulster County accident data contained relatively few observations (approximately 907 two-car accidents, and 865 single-car accidents), a test of the Koornstra Model on a large accident data base for 1980 and 1981 was selected. The analysis of the performance of the Koornstra Model on these data is presented in Section 4.2.

4.1 ULSTER COUNTY TEST

The Thorpe model and the related one- and two-car Koornstra Models are tested on the Ulster County data first. Then the more important test of the Basic Koornstra Model is presented.

4.1.1 THE THORPE MODEL AND THE ONE- AND TWO-CAR KOORNSTRA MODEL APPLIED TO THE ULSTER COUNTY DATA

The data used in all the Ulster County tests in this report are shown in Tables 4-1 and 4-2. Table 4-1 shows the accident involvement matrix for two car collisions where both cars had drivers. The driver classes are age by sex with 3 age categories: 16-24, 25-50, and 51 and up. Table 4-2 shows a breakdown of single car accidents including the case of striking a parked car.

The Thorpe Model is applied first and the resulting relative exposure estimates (normalized to class 2, males 25-50, having an estimate of 1.00) are shown in

Table 4-3. A corresponding proneness measure is obtained by dividing single car accidents by estimated esposure and as will be done repeatedly in this report normalizing to an estimate of 1.00 for males 25-50. The results are also shown in Table 4-3.

It is suggested that overall these exposure and proneness values do not accord very well with intuition. For example, in the exposure estimates, young females are estimated to have twice the exposure of young males.

The next model to be applied is the One- and Two-car Koornstra Model. Proneness and exposure estimates are given in Table 4-4.* Again, it is suggested that the estimates do not accord overall very well with intuition. In fact, the exposure estimates (except for the first category - male under 25) agree rather well with those of the Thorpe Model. Since the Thorpe model bases proneness on single car accidents, it is not expected that the proneness values should agree.** The agreement of the exposure values suggests that the One- and Two-car Koornstra model is dominated by the comparison between one- and two-car accidents on which the Thorpe model is based. It appears that the One- and Two-car Koornstra model is subject to the same criticisms as the Thorpe model.

The chi square value and the corresponding degrees of freedom shown in Table 4-4 show that there is a very significant lack of fit. This is due largely to the lack of agreement between the One- and Two-car aspects of the model as will be seen when the Basic Koornstra Model (i.e. the Two-car only model) is discussed.

In agreement with Koornstra's observation, it is concluded that the One- and Two-car Koornstra model does not fit the data very well.

*Both Maximum likelihood and least squares estimates are given (See Reference 5 for calculation procedures) to demonstrate the close agreement of the two estimates.

**Prior studies have shown that proneness is different for single and multiple car accidents, at least for older drivers.

4.1.2 THE BASIC KOORNSTRA MODEL APPLIED TO THE ULSTER COUNTY DATA

Proneness and Exposure Estimates

The results of applying the Basic Koornstra Model (i.e. the Two-car only model) to the aggregate Ulster County data are summarized in Table 4-5. The analysis of this case and disaggregations of it are the primary aim of this section.

The exposures and pronenesses as shown in Table 4-5 probably agree slightly better with intuition than those obtained for the Thorpe and One- and Two-car Koornstra Models. However, they do not appear to be entirely satisfactory. For example, females 25 to 50 have a much higher proneness than older or younger females.

Analysis of the Fit of the Basic Koornstra Model

Although there are moderately large counts in the cells of the accident matrix, a closer examination of the data suggests that the sample was not large enough to adequately test the two-car collsion model. This appears to be due to the rather subtle dependence of the proneness estimates on the accident matrix.

Consider the simple multiplicative model $\hat{X}_{ij} = W_i W_j$. This model could arise if all the proneness values in the two-car model were equal (they could then all be set equal to 1/2) so that $\hat{X}_{ij} = e_i e_j$. It could also arise from a variant of the generalized two-car model $\hat{X}_{ij} = p_i e_i p_j e_j$. In the latter case $p_i e_i$ cannot be separated into a proneness p_i and an exposure e_i. In any case, if the simple multiplicative model is postulated to hold, then no proneness estimates which vary by driver class are possible. The simple multiplicative model is easily fit to the data as noted in Section 3.6.

The expression for X^2 ("chi square") is also given in Section 3.6. For the data in Table 4.5, $X^2 = 13.7$ for the Simple Multiplicative Model. The degrees of freedom (i.e. independent parameters) in the model are six. There are 21

degrees of freedom in the data. The Basic Koornstra model has 5 more degrees of freedom (i.e. 5 more independent parameters) than the simple multiplicative model (i.e. 11-6) but reduces the chi square value only by 3.7 (i.e. 13.7 - 10.02). The change in chi square 3.7, can be referred to a chi square table with the appropriate degrees of freedom, five. The decrease in chi square is completely insignificant.

This means that the Basic Koornstra Model does not describe the data significantly better than the Simple Multiplicative Model. It can be concluded that the Basic Koornstra Model cannot be properly applied to this data at this level of aggregation*. This conclusion is also reached by examining the standard errors of the proneness values estimated in the Basic Koornstra Model. These standard errors are estimated in Appendix B.

To summarize the evidence for the fit of the Basic Koornstra Model to the aggregate accident involvement matrix the observed facts are these:

1. The X^2 of the Basic Koornstra Model is 10.0 for 11 degrees of freedom - certainly no evidence here of lack of fit.

2. However, the Simple Multiplicative Model has a X^2 of 13.7 with 15 degrees of freedom so it does not fit significantly worse and so is to be preferred as the simpler model.

3. The conclusion that the improved fit realized by the use of the Basic Koornstra Model over the Simple Multiplicative Model is not sufficient to allow for separation of exposure and proneness values. This is confirmed by the large standard errors calculated for the proneness and exposure estimates obtained from the Koornstra Model.

*It may be noticed that the chi square goodness of fit statistic for the Basic Koornstra Model, 10.02 shows no lack of fit for 11 degrees of freedom. The point is that the simple multiplicative model with much fewer degrees of freedom shows no lack of fit and does not fit significantly worse than the Basic Koornstra Model.

The Basic Koornstra Model has lead to unstable estimates of proneness and exposure. The possibility that the Basic Koornstra Model would lead to stable estimates when applied to a data set containing many more accidents may still be considered. That possibility lead to the analysis of the North Carolina data in Section 4.5.

Although a larger sample size is necessary to see if the Basic Koornstra Model does or does not fit better than the Simple Multiplicative Model in this situation, it is possible that a combination of variable proneness and incomplete mixing is at work leading to a cancellation of the tendency for a negative second eigen value due to variable proneness by a tendency for a positive second eigen value due to incomplete mixing, and hence leading to the neutral condition of the second eigen value being not significantly different from zero.

To test for the presence of this effect, a stratification of the data was performed in an attempt to eliminate the incomplete mixing through the choice of space/time intervals where the driver exposure to situations is proportional to total driving within the strata. Several stratifications of the data by time and roadway category were considered. Five different space/time strata were selected as likely to alleviate the incomplete mixing problem. These strata were as follows:

1. All accidents on State Highways

2. All accidents except those which happened late at night (8 p.m. to 5 a.m.).

3. All accidents in the jurisdiction of the Town of Kingston.

4. All accidents outside of Kingston.

5. All accidents during morning or evening rush hours.

The accident involvement matrix was computed for each of these strata. The resulting X^2 values for the Basic Koornstra Model and the Simple Multiplicative Model in each case are given in Table 4-6. In each case, except rush hour accidents, the Basic Koornstra Model did not fit significantly better than the Simple Multiplicative Model. In the case of rush hour accidents, the X^2 difference was 10.1 (5 degrees of freedom). This is significant at the .1 level but not at the .05 level. Furthermore, the probability of getting a X^2 larger than 10.1 with 5 degrees of freedom once in 5 independent tries is .3. This suggests that this result is not significant. Also, as expected, there were high standard errors on the exposure and proneness estimates. Consequently, it is concluded that the separate analysis of the five strata does not change the conclusion that the accident matrix cannot produce estimates of proneness and exposure in Ulster County.

The Basic Koornstra Model did not fit significantly better than the Simple Multiplicative Model even when the accident involvements were stratified into time/location strata which eliminated some of the incomplete mixing problems. A larger sample size might change that result and that possibility will be addressed in Section 4.5 by analyzing the larger North Carolina data set. It is also possible that the failure in the Ulster test is due largely to remaining incomplete mixing.*

*More specifically, variable proneness could be present, but not show up reliably in the model because of incomplete mixing.

Table 4-1 Two Car Accident Involvements (Ulster County)

DRIVER GROUP	MALE 16-24	MALE 25-50	MALE 51+	FEMALE 16-24	FEMALE 25-50	FEMALE 51+
Male:						
16-24	72	91	53	50	71	26
25-50		88	70	50	90	34
51+			42	32	43	28
Female:						
16-24				38	45	25
25-50					62	34
51+						28

Table 4-2 Single Car Accidents (Ulster County)

MALES:			FEMALES:		
16-24	25-50	51+	16-24	25-50	51+
297	228	74	113	104	49

Table 4-3 Thorpe Model Proneness and Exposure Estimates on Ulster County Data

	Exposure	Proneness
Male:		
16-24	.28	4.65
25-50	1.00	1.00
51+	1.04	.31
Female:		
16-24	.66	.75
25-50	1.28	.36
51+	.67	.32

Table 4-4 One- and Two-Car Koonstra Model On Ulster County Data

	MAXIMUM LIKELIHOOD	
	exposure	proneness
M16-24	.555	2.148
25-50	1.000	1.000
50+	.977	.358
F16-24	.655	.762
25-50	1.239	.384
50+	.557	.470
$X^2 = 30.26$		DF = 15
$G^2 = 29.23$		

Table 4-5 Basic Koonstra Model:
Ulster County Test Two-Car Accidents

$\lambda_1 = 326.3$

$\lambda_2 = -23.6$

$\lambda_3 = 18.5$

$\lambda_4 = 11.5$

$\lambda_5 = 8.1$

$\lambda_6 = 5.3$

	exposure (e_i)	proneness (p_i)
M16-24	.603	2.356
25-50	1.000	1.000
51+	.335	3.860
F16-24	.431	2.014
25-50	.409	4.175
51+	.342	1.674

(All estimates normalized to M25-50 = 1)

Basic Koonstra Model:

$X^2 = 10.02$ $G^2 = 9.57$ DF = 10

Simple Multiplicative Model:

$X^2 = 13.71$ DF = 15

Table 4-6 Chi Square Values for the Simple Multiplicative and Basic Koornstra Models

MODEL STRATUM	SIMPLE MULTIPLICATIVE	BASIC KOORNSTRA
1. State Highway	14.76	10.50
2. No Late Night	15.88	11.07
3. Kingston	12.68	11.73
4. Not Kingston	12.86	7.97
5. Rush Hour	15.00	4.87

4.1.3 COMPARISON OF DIRECT EXPOSURE ESTIMATES WITH INDUCED EXPOSURE ESTIMATES ON ULSTER COUNTY DATA

As noted at the beginning of this section, one reason for selecting Ulster County as the source of data to test induced exposure models was the fact that direct exposure data was to be obtained by roadside observation. When this data had been collected, it was not deemed entirely satisfactory for these purposes since its accuracy, when broken down by age and sex, was in doubt. In view of the fact that the induced exposure models did not produce acceptable exposure estimates, the adequacy of the direct exposure data was not of much importance for determining the adequacy of induced exposure models. Disagreement or even agreement of the separate exposure estimates could be due to inaccuracy in either or in both. Nevertheless, since the direct exposure data was available, a comparison could be made fairly easily and is reported in this section. The conclusion, as expected, is that the induced exposure and direct exposure show an unsatisfactorily low degree of correlation.

Hans Joksch of the Center for Environment and Man (CEM) has communicated to TSC some Ulster County exposure estimates obtained by direct roadside observations. Dr. Joksch has cautioned that the data may not be accurate enough for some purposes, especially as regards any age breakdown. Keeping in mind the possibility that disagreement (or even agreement) of the direct exposure estimates obtained from CEM with induced exposure estimates may be due to inaccuracies in either, a comparison is nevertheless reported.

The Ulster County direct exposure estimates are of two kinds: "segment VMT" and "intersection VMT." For the purposes of comparing with induced exposure estimates, they may both be considered to be estimates of total VMT which when converted to relative estimates are comparable to the induced exposure estimates.

Segment and intersection VMT estimates are shown in Table 4.8 (no late night)* and Table 4.9 (total day and night).* These are shown as relative exposure estimates in Table 4.10 (no late night) and Table 4.11 (total day and night). The corresponding induced exposure estimates using the Basic Koornstra Model are also shown for comparison in Tables 4.10 and 4.11. Relative estimates are normalized to 1.0 for males in the middle age group).

It may be seen that the induced exposure estimates do not agree very well with either the "segment" or the "intersection" estimates. Although the agreement could have been worse, where there is agreement, it could be due to the fact that accidents (on which induced exposure is based) increase in general with exposure.**

The accident rate estimates derived from the exposure estimates divide out that effect and therefore offer a more decisive comparison. Table 4.12 shows the relative (two car) accident involvement rates based on each of the three exposure estimates in Table 4.11 (total day and night). The accident rate in each case is proportional to the number of two-car accident involvements divided by the corresponding exposure. The accident rates are then normalized so that the rate for males in the middle age range is 1.0. It may be observed that the relative accident rates derived from the direct exposure estimates do not agree or even correlate well with those derived from the induced exposure (Basic Koornstra Model).

*The CEM time periods were 7-19 and 19-23. Consequently "no late night," in reference to the CEM data, refers to 7-19, while total refers to 7-23. In the case of the accident (induced exposure) data "no late night" means 6-19 while total means all times. These are not very different except for the exclusion of 23-7 in the CEM data.

**It should be noted that accidents always show some correlation with exposure and in this sense, pure accident data is "induced exposure". However, it is entirely unsatisfactory for the analytic purposes for which exposure is intended. For that a much higher degree of correlation with direct exposure is needed.

It should also be pointed out that the direct exposure estimated accident rates contradict (to the extent they are accurate) the Thorpe hypothesis. The accident rates derived from the direct exposure estimates are largest for the oldest age groups and smallest for the youngest age groups. Therefore although these data do not confirm the Basic Koornstra model they seem much more severely negative in relation to the Thorpe Model or the One and Two Car Koornstra Models (which showed, as expected, much higher proneness values for younger drivers than for older).

However, it must be kept in mind that the accuracy of the direct exposure data with respect to age is in question. Hans Joksch indicated that 25 years and 50* years had been intended as the breakpoints for the age groups but feels that both actual breakpoints were higher.

In general these data do not contradict the main findings of the study of induced exposure models on the Ulster County data, namely that the data do not support the Koornstra model but have an insufficient sample size to conclusively invalidate the model.

4.2 NORTH CAROLINA ANALYSIS

The Ulster County data was inconclusive on the usefulness of the Koornstra Model because of insufficient sample size. As far as they went, the conclusions were negative. In order to see whether a larger sample size offered a chance for a more favorable test of the Koornstra Model, a large accident data set would be useful. Such data was readily available from the University of North Carolina's Highway Safety Research Center.

*25 and 51 are the breakpoints for the accident data.

Table 4.8

NO LATE NIGHT

	SEGMENT	INTERSECTION
MALE		
Young	41853	25860
Middle	51821	33824
Old	8338	11775
FEMALE		
Young	25970	18931
Middle	28466	21492
Old	5052	6017

(for VMT multiply by 1000)

Table 4.9

TOTAL EXPOSURE

	SEGMENT	INTERSECTION
MALE		
Young	42435	32580
Middle	52286	36421
Old	8394	12851
FEMALE		
Young	26221	23248
Middle	28577	23205
Old	5066	6717

(for VMT Multiply by 1000)

Table 4.10

RELATIVE EXPOSURE - NO LATE NIGHT

	SEGMENT	INTERSECTION	BASIC KOORNSTRA
MALE			
Young	.808	.765	.585
Middle	1.000	1.000	1.000
Old	.161	.348	.358
FEMALE			
Young	.501	.560	.435
Middle	.549	.635	.449
Old	.097	.178	.395

Table 4.11

RELATIVE EXPOSURE - TOTAL

	SEGMENT	INTERSECTION	BASIC KOORNSTRA
MALE			
Young	.812	.895	.603
Middle	1.000	1.000	1.000
Old	.161	.353	.335
FEMALE			
Young	.501	.638	.431
Middle	.547	.637	.409
Old	.097	.184	.342

Table 4.12

ACCIDENT RATES (TWO CAR) - TOTAL

	SEGMENT	INTERSECTION	BASIC KOORNSTRA
MALE			
Young	1.057	.959	1.424
Middle	1.000	1.000	1.000
Old	3.940	1.797	1.894
FEMALE			
Young	1.133	.890	1.317
Middle	1.489	1.279	1.992
Old	4.270	2.251	1.211

They supplied information on all two-car accidents occuring in the State of North Carolina during 1980 and 1981 (over 100,000 such accidents in all). A table of two car accident involvements was constructed using the age-sex categories shown in Table 4.13. Besides the table representing all two-car accidents, tables representing various stratifications by time and highway, were supplied by HSRC at TSC's request. The stratification categories involved are shown in Table 4.14. There are 24 strata representing four locations by six time periods. This represents a much finer degree of stratification than was attempted with the Ulster County data. It represents a rather fine stratification consistent with the requirement of a goodly sample size in each segment for statistical significance.

When the Basic Koornstra Model was fit to the overall table of two-car involvements, it was found that the maximum likelihood fit was identical to the Simple Multiplicative Model. In other words, the maximum likelihood Koornstra Model degenerated into the special case where all proneness values are equal - the Simple Multiplicative Model. This phenomenon was not surprising since the second eigenvalue (λ_2) was positive. In Section 3.6, it was noted that the degenerate maximum likelihood solution might be expected when λ_2 is positive. It was expected on the basis of the Ulster County experience that the unstratified accident involvement matrix would lead to a Koornstra Model not significantly different from the Simple Multiplicative Model.

As a consequence of negative results of the test of the Koornstra Model on the full unstratified accident involvement matrix, the accident involvement matrices for the stratifications by the categories shown in Table 4.14 were considered.

There were 24 such matrices and many of them contained empty cells. Since there were ten matrices which did not contain any empty cells, it was decided to base the remainder of the tests on these ten accident involvement matrices.

These ten cases are represented in Table 4.15. The first column in that table indicates the case. The second column, the chi square value corresponding to the Basic Koornstra Model, and the third column, the chi square value corresponding to the Simple Multiplicative Model.

The fourth column shows the difference in chi square values. In each case, this should be referred to 10 degrees of freedom. The largest chi square (change) value is 13.3. A value larger than 13.3 occurring from a chi square distribution with 10 degrees of freedom has a probability of over .2. The probability of the largest of 10 values with chi square distributions with 10 degrees of freedom being greater than 13.3 is obviously much larger (greater than .9 if they are independent). Consequently the reduction in chi square from using the Basic Koornstra Model over the Simple Multiplicative Model is completely insignificant. There is not even a small indication that the Basic Koornstra Model fits better. It is concluded that the Basic Koornstra Model does not appear to provide an appropriate framework for analyzing the North Carolina data. Although different stratifications for the driver vehicle groups and of the overall data could have been tried, the negative results in all eleven cases (one overall, ten subsets) tried suggest that the model cannot be relied on for exposure estimates.

Table 4.13

TEN AGE-SEX GROUPS

	SEX	
AGE	MALE	FEMALE
16-20	1	2
21-25	3	4
26-45	5	6
46-65	7	8
65+	9	10

Table 4.14

TIME-LOCATION STRATA

Time Strata (time of day - day of week)

	FRI	SAT	SUN	MON	TUE	WED	THR	FRI
6 a.m. to		6	6	2	2	2	2	2
9 a.m. to		6	6	1	1	1	1	1
3:30 p.m. to		6	6	2	2	2	2	2
6:30 p.m. to	5	5	4	4	4	4	4	
9:30 p.m. to 6:00 a.m.	5	5	3	3	3	3	3	

Location Strata (highway type)

	divided	undivided
Urban	1	2
Rural	3	4

Table 4.15

COMPARISON OF BASIC KOORNSTRA MODEL WITH SIMPLE MULTIPLICATIVE MODEL ON STRATA OF NORTH CAROLINA ACCIDENT DATA

	Chi Square for Basic Koornstra Model	Chi Square for Simple Multiplicative Model	Difference in Chi Square
UD1*	44.64	54.24	10.40
UD2	47.72	56.16	8.74
UD6	31.32	37.73	6.41
UN1	109.72	118.09	8.37
UN2	53.39	60.96	7.57
UN5	86.67	94.63	7.96
UN6	49.00	59.79	10.79
RN1	43.63	53.30	9.67
RN2	44.71	57.11	12.40
RN6	44.01	57.31	13.30

*U= urban, R= rural, D= divided highway, N= non-divided highway, 1, 2,..., 6 = time periods (as defined in Table 2, e.g., UD1= urban, divided highway during time period 1)

5.0 CONCLUSIONS

The Ulster Couty accident data did not support the Basic Koornstra Model. This conclusion is based primarily on the fact that the Basic Koornstra Model did not fit the data signifcantly better than the Simple Multiplicative Model. Consistent with this, the comparison of proneness estimates was largely meaningless because of high standard errors in ratios of proneness estimates. When the data were disaggregated spatially and temporally, the problem remained i.e., the Basic Koornstra Model still did not fit significantly better than the Simple Multiplicative Model.

This test was negative on the applicability of the Basic Koornstra Model, but it was decided that a larger data set would allow a more conclusive test. The North Carolina data provided this test. The North Carolina data represented over 100,000 accidents (vs. 907 for the Ulster County data). The aggregate data showed no evidence that the Basic Koornstra Model fit better than the Simple Multiplicative Model and the data disaggregated by four highway classes and six time periods showed no evidence in support of the Basic Koornstra Model. It must be concluded that the Basic Koornstra Model is not appropriate for the North Carolina data at least for the driver categories and disaggregation categories tested.

The Thorpe and One- and Two-Car Koornstra Models were also considered but were given a much briefer treatment because of previous evidence leading to the conclusion that they have less potential for validity than the Basic Koornstra Model. The evidence given here supports that conclusion.

It is concluded that in all probability, the Koornstra type and Thorpe type induced exposure models are not suitable for deriving driving exposure from highway accident data. It appears that these models give grossly inaccurate, unreliable and/or inconsistent results.

Appendix A

The Aggregation Theorem for Koornstra-Type Models

Let there be M classes of drivers labelled 1, ..., M and let the accident involvement matrix between these driver classes be described by a Generalized Koornstra Model:

$$X_{kj} = (p_k + p_j + \alpha\ p_k p_j + \beta)\ e_k\ e_j \qquad k,j = 1..., M$$

If these classes are aggregated into a new set of classes K, J etc., then the new accident involvement matrix is given by:

$$X_{KJ} = \sum_{k \subset K} \sum_{j \subset J} X_{kj}$$

Here the notation $k \subset K$ means that the aggregate class K contains the drivers in the class labelled by k. Of course only one aggregated class contains drivers in the class labelled by k.

The above expression for X_{KJ} results from the fact that X_{KJ} stands for the number of involvements of drivers in class K in collisions with drivers in class J.

To simplify the derivation let $t_k = p_k e_k$ then

$$X_{kj} = t_k e_j + e_k t_j + \alpha\ t_k t_j + \beta\ e_k e_j$$

Then

$$X_{KJ} = \sum_{k \subset K} \sum_{j \subset J} X_{kj} = \sum_{k \subset K} \sum_{j \subset J} (t_k e_j + e_k t_j + \alpha\ t_k t_j + \beta e_k e_j)$$

$$= T_K E_J + E_K T_J + \alpha\ T_K T_J + \beta\ E_K E_J$$

Where

$$E_K = \sum_{k \subset K} e_k$$

and

$$T_K = \sum_{k \subset K} t_k = \sum_{k \in K} p_k e_k$$

consequently

$$X_{KJ} = (P_K + P_J + \alpha P_K P_J + \beta) E_K E_J$$

where

$$P_K = T_K / E_K$$

Note that in particular the driver classes labelled by k could be individual drivers. In this case, X_{kj} would represent the <u>expected</u> number of collisions between drivers k and j* in the time period (consequently X_{KJ} represents the expected number of involvements of members of class K with members of class J). Since the expected number of collisions between two specific drivers is very low, X_{kj} also would represent the <u>probability</u> of a collision in that case (as noted in Section 3.5 Xkk = 0 but that presents no problem since only a negligible fraction of a driver's collisions are with a specific other driver, in any case).

Note that the aggregation theorem as derived for the Generalized Koornstra Model specializes to the Basic Koornstra Model by taking $\alpha = \beta = 0$.

*Since k now represents a single driver rather than a class of drivers, in keeping with standard notation, in this case, one replaces "$k \subset K$" by "$k \in K$" in e.g. the definition of e_K.

Appendix B

Variance Estimates for Comparisons of Exposure and Proneness Estimates

In Section 4.1, it is concluded for the Ulster County Data that the Basic Koonstra Model does not describe the data significantly better than the Simple Multiplicative Model. The sample size was too small for an adequate test of the Basic Koonstra Model.

The same conclusion can be reached when the standard errors in the parameters of the Basic Koornstra Model are calculated. It is not to be expected that the standard errors in the exposure estimate are exceedingly large but is is to be expected that comparisons of proneness values are invalidated by large standard errors in these comparisons.

Standard errors in model parameters for the Basic Koornstra Model can be calculated from the Fisher information matrix. The calculation (as described in Reference 5) is based on the assumptions of model validity, and large sample size which are themselves in question. However, they provide the only estimates available and if they indicate large errors, the errors in the parameters are almost certainly correspondingly large. These variances and covariances in model parameters lead directly to estimates of variances in comparisons of exposure and proneness estimates. The only assumption needed is that the mean of each model parameter be substantially greater than its standard error. Again, a failure in these assumptions is not expected to nullify the validity of any observation of high variances.

The resulting variance estimates are shown in Table B-1. This table shows the square of the estimated fractional standard error (i.e. it shows the variance ratio) in the estimates of e_i/e_j and p_i/p_j. Thus if $R = e_i/e_j$ or $R = p_i/p_j$ then the table shows σ^2_R / R^2. For example, the estimate of p_5/p_1 from Table 4.5 is 1.77. Since the i = 1, j = 5 entry Table B-1 is .216, the estimate of the variance ratio of this estimate is .216. When a proportionate standard error is large (nearly 1 or even greater than 1) it is sometimes useful to consider it as a

standard error in the logarithm (to the base e) of the estimate. In other words, the estimate is of σ for $\log_e R$. Crudely, a 95 percent confidence interval on $\log_e R$ is $\log_e R \pm 2\sigma_R/R$.

For example,

$\log (P_5/P_1) = \log (1.77) \pm 2 \sqrt{.216}$

or

$$.7 \leq P_5/P_1 \leq 4.5$$

gives a very crude 95 percent confidence interval.

Table B-1 Variance Ratios for Exposure and Proneness Comparison

$\sigma^2_{R/R}$

ij	R_e	R_p
1,2	.040	.193
1,3	.074	.213
1,4	.065	.250
1,5	.076	.216
1,6	.034	.130
2,3	.040	.129
2,4	.029	.138
2,5	.035	.110
2,6	.030	.167
3,4	.053	.158
3,5	.073	.168
3,6	.064	.210
4,5	.056	.167
4,6	.056	.247
5,6	.058	.188

$R_p = P_i/P_j$

$R_e = e_i/e_j$

ij = 1,6

1 = Male 16-24
2 = Male 25-50
3 = Male 51+
4 = Female 16-24
5 = Female 25-50
6 = Female 51+

Appendix C

Number of Encounters Between Vehicles Moving at Different Speeds

Suppose that N_i vehicles of type i travel e_i miles per unit time (this is exposure, not speed) uniformly over an entire two-lane roadway system, in both directions. While on the roadway they travel at speed V_i. The total number of miles of the roadway lanes is L. Then the number of expected encounters per unit time between vehicles of type 1 and type 2 is

1. $N_i e_i \; N_2 \, e_2 \; (1/V_1 + 1/V_2)/L$

   if vehicles of type 1 and type 2 are going in opposite directions.

2. $N_1 \quad e_1 \quad N_2 e_2 \quad (1/V_1 - 1/V_2) \, /L$

   if vehicles of type 1 and type 2 are going in the same direction.

3. $DN_1 e_1 \; N_2 e_2 \; (1/V_1 + 1/V_2)/L^2$

   for encounters at intersections where D/L is the fraction of roadway length in intersections.

The main question which this appendix seeks to address is whether the number of encounters can be split into the product of two factors as required by assumption 1 in the derivation of the Basic Koornstra Model. Clearly these expressions do not factor in the required manner.

In the aggregate with a combination of same way, different way and crossing encounters being present, and with realistically each vehicle travelling at various speeds, it is not clear how seriously the number of encounters fails to factor but it appears to be potentially serious.

The remainder of this Appendix is devoted to a brief outline of the derivation of the formulas for the number of encounters.

Each car of type i travels e_i miles per unit time over a system of length L. This means each car enters the system e_i/L times per unit time. The time to traverse the system is L/V_i. Therefore, the total numberof vehicles of type i in the system at any time is $N_i(e_i/L)\,(L/V_i) = N_ie_i/V_i = n_i$.

The number of vehicles of type i per unit length is n_i/L.

The number of vehicles of type 1 per unit time encountered by a vehicle of type 2 going in the opposite direction is thus $n_1(V_1 + V_2)/L$. Therefore, the total number of encounters between vehicles of type 1 with vehicles of type 2 going in the opposite direction is $(n_1/L)\,(V_1 + V_2)\,n_2 = N_1e_1\ N_2e_2\ (1/V_1 + 1/V_2)/L$. This proves the first formula - for encounters of vehicles going in different direction. The second formula - for encounters going in the opposite direction is derived similarly. The third formula - for encounters at crossings is derived a little differently.

Each car of type 1 enters a given intersection e_1/L times per unit time. Each time it enters the intersection, it spends d/V_1 units of time in the interstetion (here d is the width of the intersection). The number of vehicles of type 2 which enter the intersection during this time going in a cross direction is $(N_2d/V_1)\,(e_2/L)$. Thus the total number of cross intersection encounters of vehicles of type 1 with vehicles of type 2 where the vehicle of type 1 enters the intersection first is $(N_1e_1/L)\,(N_2d/V_1)\,(e_2/L) = (N_1e_1\ N_2e_2/V_1)\,(d/L^2)$.

The total number of such encounters where the vehicle of type 2 enters the intersection first is $N_1e_1\ N_2e_2/V_2\ \ d/L^2$.

Therefore, the total number of intersection encounter regardless of which enters first, is $(N_1e_1\ N_2e_2\ (V_1V_2))\,(1/V_1 + 1/V_2)\ \ d/L^2$.

If we sum this up over all intersections, we get

$$N_ie_i\ N_2e_2\ (1/V_1 + 1/V_2)/V_iV_2\ \ D/L^2$$

where $D = \sum d$ is the sum of the width of all the intersections and $D/L = f$ is the fraction of the roadway in intersections.

REFERENCES

1. "Literature Review of Induced Exposure Models," Project Memorandum; PM-223-05-4A. P. Mengert (June 1982).

2. "Calculating Relative Involvement Rates in Accidents Without Determining Exposure," Traffic Safety Research Review. John D. Thorpe (March 1967).

3. "A Model for Estimation of Collective Exposure and Proneness from Accident Data," Accident Analysis and Prevention, Volume 5 (1973), Pergamon Press. Matthys J. Koornstra, and "Empirical Results on the Exposure-Proneness Model," Accident Analysis and Prevention, Volume 5 (1973), Pergamon Press. Matthys J. Koornstra.

4. Traffic Accident Exposure and Liability. Carsten Wass, Allerod, Denmark (1977). Printed in Denmark by ROC, Rungsted (privately published, has been available from author at: Ligustervangen 57, DK 3450, Allerod).

5. "Approach to Testing Induced Exposure Models," Project Memorandum; PM-223-05-7. P. Mengert (August 1982).

6. "Driver Exposure - The Indirect Approach for Obtaining Relative Measures," DOT-HS-820, 179. Ezio C. Cerelli (March 1972). Also a shortened version in Accident Analysis and Prevention, Volume 5 (1973).

7. "Induced Exposure," Accident Analysis and Prevention, Volume 5. Frank Haight (1973).

8. "A Statistical Analysis of Rural Ontario Traffic Accidents Using Induced Exposure Data," Accident Analysis and Prevention, Volume 1. Brian Carr (1969).

9. "An Empirical Analysis of Accident Data Using Induced Exposure," HIT Lab Report. William K. Hall, (September 1970).

10. Computational Mathematics. B.P. Demidovich and I.A. Maron, MIR Publishers, Moscow, U.S.S.R. (English translation, 1981).

11. "Measuring Exposure," by Diccon Bancroft and John Riemersma in Contingency Table Analysis for Road Safety Studies edited by Gerald A. Fleischer, Sytoff and Nordoff, The Netherlands (1981).