WHY ARE EXPERTS NOT BETTER IN JUDGING THE DANGER OF FILMED TRAFFIC CONFLICTS?

HERMAN W. KRUYSSE
Unit of Experimental and Theoretical Psychology/Centre for Safety Research, Leiden University, The Netherlands

and

GERT JAN WIJLHUIZEN
Institute of Preventive Health Care, NIPG-TNO, Leiden, The Netherlands

(Received 6 August 1990)

Abstract—Earlier studies on the subject of subjective judgment of traffic conflicts showed that untrained subjects can reliably judge the dangerousness of filmed traffic conflicts. It was concluded that these judgments were based on a concept of dangerousness. In line with these findings two experiments were conducted. Experiment 1 tested the hypothesis that traffic experts (traffic engineers trained to evaluate and improve traffic safety) and lay people use the same concept of dangerousness when judging filmed conflicts. Experiment 2 investigated which aspects of the conflicts are considered by experts when making these judgements. The results show that (i) experts and lay people are equally reliable in judging traffic conflicts, (ii) experts base their judgement on the same concept of dangerousness, (iii) experts do not base their judgement on aspects that they themselves regard as important for the conflict, (iv) a weak relation is found between violations of traffic regulations and judgements of dangerousness.

INTRODUCTION

Traffic conflicts are considered as indices for traffic safety. A traffic conflict is a near-accident—a dangerous situation in which two or more road users approach each other to such an extent that there is a risk of collision. It is supposed that traffic conflicts are related to accidents in such a way that they reflect traffic safety and provide diagnoses of safety problems at particular locations. Thus the registration of traffic conflicts quantifies traffic safety.

However, it appeared that traffic conflict techniques have low validity (Williams 1981; Engel 1985). The reason mentioned is that most of these techniques were based on poor definitions of what a conflict is and on subjective judgements; unreliable registration of conflicts was considered as a significant cause for the low validity.

As a consequence researchers turned their attention towards objective conflict techniques. However the definition of objective technique is confusing, since it includes techniques based on an objective definition of the conflict but not on objective registration. Objective definitions of conflicts are always based on the time dimension; Hayward's (1972) time to collision (TTC), Hydén's (1975) time to accident (TTA), Balasha, Hakkert, and Livneh's (1980) definition incorporating both lateral and longitudinal decelerations and accelerations. Objective analysis of conflicts is performed by a frame-by-frame analysis of filmed conflicts (van der Horst 1982). Contrary to objective analysis of filmed conflicts, in general, human observers are used for the subjective judgment of traffic conflicts in the field.

Literature showed that reliability is at stake when subjects are used to estimate objectively defined conflicts (Shinar 1984; Kruysse and Wijlhuizen 1988; Kruysse 1990). These studies showed that untrained subjects are reliable in judging the dangerousness of filmed traffic conflicts even when the notion conflict is not introduced at all. It was shown that these judgements are based on an internal concept of dangerousness. The significance of these findings was explained in terms of construct validity (Kruysse 1990).

The significance of the internal concept becomes clearer, however, when judgements of untrained observers are compared with judgements of traffic experts. For traffic experts claim to improve traffic safety by evaluating conflicts at dangerous spots. We
defined experts as traffic engineers, employed by municipal works and trained to evaluate and improve traffic safety. In a short questionnaire it was found that no general standardized observation techniques are used by traffic experts allied to sections of municipal works in many Dutch cities. Traffic situations are subjectively judged on the spot or from a videoscreen, apparently with the guidance of informal techniques.

This situation is elaborated upon in this study. The study contains two experiments. Experiment 1 was designed to, first, investigate the interrater and intrarater reliability of traffic experts. Fifteen traffic experts from four main Dutch cities were instructed to judge the dangerousness of filmed traffic conflicts. In earlier studies, the same conflicts were judged, under equal conditions, by untrained observers (Kruysse and Wijlhuizen 1988; Kruysse 1990). Based on the results of these studies and the assumption that trained experts should be highly reliable, we hypothesized that experts are more reliable in judging conflicts than untrained observers. Second, we tested the hypothesis that experts and untrained observers have the same concept of dangerousness. This is obviously a desirable feature for the application of conflict observation techniques.

Given the existence of a concept of dangerousness, it is still not clear on what aspects of conflicts the judgements are founded. This was investigated in Experiment 2. Four months after Experiment 1, the same experts were asked to evaluate the same filmed conflicts, by pointing out noteworthy aspects.

**EXPERIMENT 1: METHOD**

*Method*

Apparatus, materials, and design were equal to those of earlier studies on this topic (Kruysse and Wijlhuizen 1988; Kruysse 1990).

*Apparatus*

Ten monochrome video monitors were connected to one U-matic (SONY) video tape recorder.

*Materials*

A total of 59 different video-taped traffic conflicts at intersections were selected. The values of the objective variable Minimal Time To Collision (MTTC) was calculated as defined by Hayward (1972) and van der Horst (1982). Only traffic conflicts with MTTCs between 2 and 0 seconds were selected and proportionally divided over the stimulus set. The set contained only conflicts with opposing traffic and traffic from the side. Conflicts were possible between all types of road users including pedestrians.

*Subjects*

A total of 15 male traffic experts working at traffic sections of municipal works in four Dutch cities—Leiden (three subjects), Rotterdam (three subjects), Utrecht (three subjects), and Amsterdam (six subjects)—participated as unpaid expert judges. All subjects had experience with conflict observation either on the spot or from video recordings. The number of years of experience among the subjects varied from 1 to 25 years, the mean and median experience was about 10 years.

*Design*

The traffic conflicts were presented in four different sessions. The first was an introductory session in which 20 different conflicts were presented. The introductory session was followed by three different experimental sessions in each of which 13 different conflicts were presented twice, so that an experimental session contained 26 conflicts. The conflicts were presented in a random sequence, with the restriction that equal presentations were separated by at least four other conflicts. Thus, the second presentation of the same conflict is always preceded by four or more other conflicts. The order of presentation of situations within each session was identical for each subject. The order
of presentation of the experimental sessions for subgroups of five subjects was according to a Latin Square design.

**Procedure**

The traffic conflicts were presented on a video monitor. Subjects were instructed to judge the dangerousness of the traffic conflicts on a 20-point scale. The extremes were named as "not dangerous" and "very dangerous." On the score form, above the scale, a simple picture of the conflict situation was given. Thus, to avoid mistakes about the intended conflict to be judged, subjects were informed about where on the intersection and between whom the conflict occurred.

**EXPERIMENT 1: RESULTS**

The responses were analysed using OVERALS (Verdegaal 1986; van der Burg, de Leeuw, and Verdegaal 1988), which performs a type of principal component analysis (PCA). OVERALS determines the extent of correspondence between (sets of) variables (i.e. the subjects). When every set in the analysis contains one variable, the OVERALS solution equals a nonlinear Principal Component Analysis, PRINCALS. The difference between a common PCA and PRINCALS is that the latter also use weights based on the level of measurement: i.e. ordinal transformations of rating scores. Thus, OVERALS transforms the judgements non-linearly into "object scores" and projects these into a multidimensional space, such that an optimal fit between the scores is generated. An object score reflects the judgements of all subjects to one object (i.e. the conflict). The transformation of judgements into object scores depends on the level of measurement of the judgements. In this experiment the rating scores were treated at an ordinal level. The optimal fit is found when dimensions added to a solution result in a minimum change of stress. The maximum value of the total fit equals the number of requested dimensions if all the variance is accounted for by these dimensions. In a three-dimensional solution the maximal fit is 3.000 when all variance in the data is explained by the dimensions. The variance accounted for by each dimension is specified by the eigenvalue.

In the first part of the analysis, OVERALS was used to determine the interrater and intrarater reliability. This was done by treating subjects as separate sets of variables and establishing a common dimension that explains a maximum proportion of the variation in the responses. The best solution of the data was given by a one-dimensional solution. One-dimensional solutions were computed for the first and subsequent judgements.

*Interrater reliability. A preliminary analysis showed that in each solution the eigenvalue of the dimension, which is labeled "subjective dangerousness," was 0.449. Thus, the dimension explained 45% of the total variance that reflects the interrater reliability. The contribution of each subject to the common dimension was reflected in the component loadings of individual subjects (Table 1).

High component loadings indicated high correlations with the dangerousness dimension, whereas low component loadings indicated low correlations. Four subjects (2, 7, 9, 11) had relatively low component loadings (see Table 1). Although this suggests

<table>
<thead>
<tr>
<th>Subject</th>
<th>Component loadings</th>
<th>Subject</th>
<th>Component loadings</th>
<th>Subject</th>
<th>Component loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.806</td>
<td>6</td>
<td>0.723</td>
<td>11</td>
<td>0.291</td>
</tr>
<tr>
<td>2</td>
<td>-0.118</td>
<td>7</td>
<td>0.346</td>
<td>12</td>
<td>0.841</td>
</tr>
<tr>
<td>3</td>
<td>0.710</td>
<td>8</td>
<td>0.791</td>
<td>13</td>
<td>0.712</td>
</tr>
<tr>
<td>4</td>
<td>0.839</td>
<td>9</td>
<td>-0.234</td>
<td>14</td>
<td>0.794</td>
</tr>
<tr>
<td>5</td>
<td>0.844</td>
<td>10</td>
<td>0.674</td>
<td>15</td>
<td>0.699</td>
</tr>
</tbody>
</table>

*The dimension is determined by the optimal scaled judgements of all conflicts; the component loading represent the correlation of an expert with the dimension.*
that these subjects had deviating opinions with respect to the dangerousness of the conflicts, the low component loadings were explained by the dichotomous use of the scale by these subjects.

As a result, 11 subjects were retained for the final analysis. The eigenvalue of a one-dimensional solution yielded 0.593. Thus the dimension explained 59% of the total variance.

**Intrarater reliability.** Each situation was judged twice during the experiment. A comparison of these two judgements is used as an indication of the consistency of the concept of dangerousness within each subject. The intrarater reliability, \( r = 0.96 \) was calculated by Pearson product moment correlation between the optimal scaled responses of the first and repeated presentation. This means that more presentations of the same conflict are judged consistently by each expert. The correlation based on the optimal scores gives the upper limit; the mean Spearman rank (S.R.) correlation based on the raw scores, \( Rs = 0.80 \), gives the lower limit.

The second part of the analysis was used to test the hypothesis that experts differ from untrained subjects in their intrarater reliability and in their judgement of conflicts.

A one dimensional OVERALS solution for untrained judges (30 students), explained 59% of the total variance (Kruysse and Wijlhuizen 1988) equal to 59% for 11 expert judges but more than 45% for all (15) experts. These values are an indication for the measure of agreement between subjects with regard to their subjective evaluation of dangerousness within each group.

Since we adapted a one dimensional representation of dangerousness, the interrater reliability can adequately be described by intercorrelations between subjects. To compare the intrarater reliability of untrained judges and experts, a Mann–Whitney test was carried out to test whether the two distributions of both sets of intercorrelations differ significantly. A significant difference was found between intercorrelations of untrained judges (\( n = 30 \)) and all experts (\( n = 15 \)); \( Z = -7.90, p < 0.01 \) (two-tailed). No significant difference was found between intercorrelations of untrained judges (\( n = 30 \)) and experts with high component loadings (\( n = 11 \); \( Z = 1.57, p = 0.12 \) (two-tailed).

To compare the judgements of lay persons with the judgements of experts, a Pearson product moment correlation, \( r = 0.86 \), between the optimal scaled responses of the two groups was calculated (Fig. 1). This means that conflicts are judged in the same way by experts and lay persons.

**DISCUSSION AND CONCLUSIONS**

The first aspect of expert judging behavior under investigation was the measure of agreement within the group of experts. A one-dimensional solution suggested that judge-
ments were based on the same concept of dangerousness, although the eigenvalue (45%) implied a moderate agreement between the judgements. Low component loadings of four experts showed that these experts deviated considerably from other experts. The difference between these four experts and the remaining 11 experts was explained by the partial, almost dichotomous, use of the scale by four experts. The agreement between judgements increased considerably in the final analysis (11 experts); the dimension dangerousness explained 59% of the total variance. The intrarater reliability $r = 0.96$, suggests that experts use consistent strategies in judging conflicts.

The second aspect of expert judging behavior under investigation was the difference between expert judgements and lay person judgements of the same conflicts. First, the difference in consistency between both groups was examined; are experts more consistent in their judgement of dangerousness than lay persons? The eigenvalues of both groups did not differ. Moreover, the distribution of intercorrelations of lay persons did not differ from the distribution of the intercorrelations of the 11 experts; this indicates that possible differences in judgements between both groups are not caused by the differences within each group. Thus, both groups are equally consistent in their use of the concept of dangerousness.

Second, the differences between the judgements of both groups was examined: will a conflict ranked by a lay person be ranked equally dangerous by an expert? As indicated by the correlation between the object scores of both groups, $r = 0.86$, untrained judges agree in general with experts in their judgement of dangerousness.

It is concluded that untrained judges and experts have the same concept of dangerousness and are equally reliable in the consistent use of this concept. This conclusion suggests that subjective conflict observations by lay persons are valid relative to experts.

This finding is important considering the uniform application of conflict observation. As mentioned earlier, however, the aspects of the conflict that give rise to these judgements remain speculative. Since experts claim to base their subjective evaluation of traffic conflicts on their knowledge concerning traffic processes, a second experiment was conducted to investigate what aspects of conflicts determine the dangerousness of these conflicts. Experts were instructed to give keywords or statements of the noteworthy aspects of each conflict, considering the specifics of the conflict. In order to relate these aspects to the judgements of dangerousness, the results of Experiment 1 and Experiment 2 were analysed using a within-subject design. Since the recency of the judgements of dangerousness in Experiment 1 could influence the evaluation, Experiment 2 was carried out four months later. Unfortunately, this time interval resulted in a drop-out of subjects; eight experts were not available for participation in Experiment 2 for personal reasons.

**EXPERIMENT 2: METHOD**

**Apparatus, materials**

Apparatus and materials were equal to those used in Experiment 1.

**Subjects**

A total of 7 male traffic experts who participated in Experiment 1—Leiden (two subjects), Rotterdam (two subjects), Utrecht (one subject) and Amsterdam (two subjects)—served as subjects.

**Design**

Three different groups of 13 different filmed traffic conflicts were presented in one session. Two orders of presentation of situations were composed by changing the order of presentation between the groups of conflicts.

**Procedure**

The traffic conflicts were presented in real time on a video monitor. Subjects were asked to evaluate each conflict situation separately. The only direction for evaluation was given by the question: Which aspects of the presented conflict situation do you consider as noteworthy in qualifying the conflict?
Subjects were allowed only to replay the conflicts in real time. They were instructed to reflect their evaluations in keywords or short statements of the noteworthy aspects of the conflicts. The total number of keywords and statements were limited to a maximum of ten. Before evaluating the next conflict, subjects were instructed to rate the importance of each generated keyword and/or statement on a 7-point scale. To avoid mistakes, simple pictures of the conflict situation were used to remind subjects about where on the intersection and between whom the conflict occurred.

RESULTS

Generally more statements than keywords were given. Subjects used three to five statements per conflict in a description and up to a maximum of seven keywords if only keywords were used. An example of statements given as a short description goes as follows:

The car driver did not yield right-of-way. His speed was too high considering the location, and therefore he took too much risk. The cyclist evaded a collision by anticipating on the car driver’s behavior, by decelerating, and by evading the car. There was no danger for other road users.

Preliminary to the analysis, the keywords and the keywords in the statements were analysed on their content: content analysis (Krippendorff 1980; Weber 1985) is a technique in which inferences are made about the text by systematically and objectively identifying specific characteristics into conceptual categories. In this study, the keywords in particular appeared a defensible criterion in identifying categories that represent noteworthy aspects of conflicts. In order to account for the validity of the categories, the descriptions were independently analysed on their content by two reviewers; \( r = 0.95 \). Different categories were discussed and in mutual agreement recategorized. A total of six main categories and six subcategories were distinguished, respectively:

WHO (category 1) refers to the directly involved road users in the conflict. The category is divided in two subcategories:
- Fast Traffic represents all motorised traffic,
- Slow Traffic represents cyclist and pedestrians.

WHAT (category 2) refers to remarks on the direct cause of the conflict. The category is divided in two subcategories:
- Risk refers to remarks on risk taking,
- Rule refers to remarks on the violation of traffic regulations.

ACTion (category 3) refers to remarks on the kind of action undertaken by one or all of the involved conflict partners. The category is divided in two subcategories:
- ANTcipation,
- MANoeuvre refers to remarks on braking and evasive actions.

THird Party (category 4) refers to remarks about road users indirectly involved in the conflict.

SPEED (category 5) refers to remarks regarding the speed of the road users.

SITuation (category 6) is a kind of rest category. The category refers to typical remarks on the location, i.e. “a bad priority regulation” and, “the situation is clearly arranged.”

From the preliminary analysis, a three-dimensional matrix was constructed: 39 conflicts × 7 experts × 7 categories. Since information about those involved in the conflict was given by the experimenter, the first category, WHO, was used only to confirm the evaluation of the intended conflicts. The cells of the matrix contained a multivariate dependent variable: rating scores of importance.

The frequency of occurrence of categories was determined by the number of “importance weights” per category over the experts and conflicts. The importance weight of each category per conflict was determined by calculating the median of the judgements.
Judging the danger of filmed traffic conflicts

Table 2. The frequency of occurrence of seven categories*

<table>
<thead>
<tr>
<th>(N) Experts</th>
<th>Risk</th>
<th>Rule</th>
<th>Ant</th>
<th>Man</th>
<th>ThP</th>
<th>Speed</th>
<th>Sit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>-</td>
<td>13</td>
<td>2</td>
<td>22</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>3</td>
<td>12</td>
<td>8</td>
<td>5</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>1</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>9</td>
<td>-</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>9</td>
<td>-</td>
<td>5</td>
<td>-</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>5</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>7</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Σ Et *Ci</td>
<td>83</td>
<td>184</td>
<td>52</td>
<td>134</td>
<td>39</td>
<td>72</td>
<td>63</td>
</tr>
</tbody>
</table>

*The first column contains the number of experts, the next columns contain the number of categories judged by the number of experts in column one. The frequency of occurrence per category is given by the sum of products of the number of judged conflicts and the number of experts.

The results are presented in Table 2. It shows the frequency of occurrence of each category judged by the number of experts per conflict; e.g. Risk was judged in seven conflicts by one single expert. Rule was judged in seven conflicts by seven experts.

In the second part of the analysis, the significance of the aspects was determined by comparing the importance weights of each category to the dangerousness scores in Experiment 1 and predicting the dangerousness on the bases of the importance weights of the aspects. A linear regression model explains 26.2% of the total variance when dangerousness is predicted from the importance weights of Rule, Sit, ThP, and Risk. The variables Speed, Man and Ant together explained 0.5%.

Since the generalizability of the judgements is dependent on the number of judgements per conflict in a category, only categories with three or more judgements per conflict are considered relevant. As a consequence, only Rule and Manoeuvre are representative (see Table 2) to act as important aspects of dangerousness; if all conflicts were scored by seven judges per category, 65% of Rule and 43% of Manoeuvre was judged by three or more experts. The importance weights of these categories were mutually compared and compared to the object scores of dangerousness. The left part of Table 3 shows the Spearman rank correlation between the importance weights of the categories and between all categories and the object scores. The category Rule explains most of the variance.

The last part of the analysis was performed to clarify the difference between the frequency of occurrence of aspects per conflict and the judged importance of the aspects; the validity of the importance weights is at stake if these weights are exchangeable for their frequency of occurrence. The right part of Table 3 shows the Spearman rank correlations.*

Table 3. Spearman rank correlations*

<table>
<thead>
<tr>
<th></th>
<th>Dang (N = 11)</th>
<th>Manoeuvre (Imp)</th>
<th>Dang (N = 11)</th>
<th>Imp × Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule</td>
<td>0.419</td>
<td>0.124</td>
<td>-0.159</td>
<td>0.071</td>
</tr>
<tr>
<td>Man</td>
<td>0.086</td>
<td></td>
<td>0.038</td>
<td>0.029</td>
</tr>
<tr>
<td>Sit</td>
<td>0.223</td>
<td></td>
<td>0.374</td>
<td>0.405</td>
</tr>
<tr>
<td>Speed</td>
<td>-0.132</td>
<td></td>
<td>0.273</td>
<td>0.568</td>
</tr>
<tr>
<td>Risk</td>
<td>-0.049</td>
<td></td>
<td>-0.051</td>
<td>0.388</td>
</tr>
<tr>
<td>Ant</td>
<td>-0.008</td>
<td></td>
<td>-0.000</td>
<td>0.468</td>
</tr>
<tr>
<td>ThP</td>
<td>-0.072</td>
<td></td>
<td>-0.089</td>
<td>0.801</td>
</tr>
</tbody>
</table>

*Column one contains the Spearman rank correlations between the scores of dangerousness and the importance weights of the categories. Column two contains the S.R. correlation between the importance weights of the categories Rule and Manoeuvre. Column three, contains the S.R. correlations between the scores of dangerousness and the frequency of occurrence of the categories. Column four contains the S.R. correlations between the importance weights and frequency of occurrence of the categories.
correlations between the frequency of occurrence and dangerousness for each category and the correlations between the frequency of occurrence and the importance weights of each category.

DISCUSSION AND CONCLUSIONS

The results of Experiment 2 reveal some insights in how traffic experts evaluate consistently conflicts. The first insight is: when asked, experts generate about nine categories of aspects that they think are important for different types of conflicts. The main categories are based on the subjects' consistent use of the following questions; who is involved, what caused the conflict, how the directly and indirectly involved road users reacted? Typically, the cause of the conflict refers frequently to remarks on the violation of traffic regulations and remarks on the type of braking and evasive action.

The second insight is: judged aspects on a scale of importance are bad predictors for the dangerousness of the conflicts. Only a fourth part of the total variance of dangerousness is explained by the importance weights when no restrictions with respect to the frequency of judgements are imposed. Thus, a discrepancy is revealed between what experts do when judging conflicts and what they say they do when judging conflicts. If experts rationally evaluate aspects of a conflict in order to qualify the conflict, one would expect the judged importance of these aspects to explain a greater part of the variance of the dangerousness. Clearly, this is not the case; only Rule explained significantly only 18% of the dangerousness.

The third insight is: frequency of occurrence of significant aspects is not exchangeable for their judged importance: Rule Freq is not related to Rule Imp and predicts dangerousness poorly (see Table 3). This suggests that when an aspect is found predicting, the rate of importance and not the frequency of occurrence represents its significance. In contrast to aspects that are found to be insignificant, the importance and the frequency of occurrence do not differ with the exception of Manoeuvre.

The fourth insight is: since violations of traffic regulations explained one-fifth of the total variance of dangerousness, this aspect suggests some importance in judging the dangerousness of conflicts.

The results of this experiment are clear. When subjects (traffic experts) are instructed to rationally qualify traffic conflicts by successively generating noteworthy aspects and judging the importance of these aspects, these aspects have a low validity relative to the dangerousness of these conflicts. Although the strategy to evaluate conflicts by asking who is involved, what is involved, and how road users are involved seems proper at face value, only the aspect concerning remarks on violations of traffic regulations are valid. Obviously, this aspect is not unique since it explains only one-fifth of the total variance of dangerousness.

In what way this aspect is related to the dangerousness of the conflict remains, however, unclear: Is the aspect concerning the violations of traffic regulations considered when experts evaluate the dangerousness of traffic conflicts? Or is this aspect found to be important by experts because they were forced to generate an aspect that seemed of relevance and unknown to them was valid relative to dangerousness? Irrespective of the answer to this question, the finding itself reveals insight into the dangerousness of conflicts. It suggests at least an indicator for dangerousness and it relates dangerousness, at least in a conceptual way, to traffic safety. Violation of traffic regulations threatens traffic safety, although the opposite statement, no violations of regulations implies safety, is invalid.

One of the problems in this kind of study is to generalize the results. This is not so much from the use of filmed conflicts; experts use that kind of material daily in their evaluations. Nor is it the validity of relating importance weights of the aspects to dangerousness, since experts claim to consider these aspects during the qualification of conflicts and since the importance of an aspect is rated relative to other aspects per conflict. Nor did the number of experts affect the results, since by definition this population is small. Nevertheless, the generalizability of the technique, employing generation and categorization of statements, could be challenged.
Judging the danger of filmed traffic conflicts

In this respect we attempted to preserve generalizability in several ways: first, the instruction to the subjects—experts were instructed to evaluate the conflicts in their normal way of duty with the exception that they were asked to mention the aspect that contributed to their judgement; second, they themselves were asked to employ the categories that would be used for later scoring; third, keywords, statements, and descriptions were sorted by two independent judges into precise categories; fourth, only categories with three or more judges per conflict were considered suitable for further analysis.

Considering the arguments above we feel that the results of these experiments are quite conclusive: (i) experts and lay persons are equally reliable when judging traffic conflicts; (ii) experts and lay persons base their judgement of traffic conflicts on the same concept of dangerousness—experts do not differ from lay persons in rating conflicts with respect to dangerousness; (iii) experts do not base their judgement of conflicts on aspects that they regard as important for that conflict; and (iv) it is unclear how the aspect concerning violation of traffic regulations is related to dangerousness—however, the fact that there is a relation is conceptually important in understanding dangerousness and its significance for traffic safety.

Acknowledgements—The authors thank Willem J. Heiser of the Dept. of Data Theory for his suggestions and comments on the analyses used in this study.

REFERENCES


Verdegaal, R. OVERALS. Leiden, The Netherlands: Department of Data Theory (UG-86-01), University of Leiden; 1986
