3.7 2nd International Traffic Conflicts Technique Workshop

May 10th - 12th, 1979

A study of observer variability and reliability in the detection and grading of traffic conflicts

A. Lightburn and C. I. Howarth

Department of Psychology,
University of Nottingham,
England.

Summary of method and major findings

After a one hour training session forty-two observers watched six films, each made up of 12 clips of 16 mm silent colour film (72 clips in all). Each clip was 25 - 30 seconds long and were separated from one another by blank pieces. The films were from three real life traffic situations. The films were shown in a different order on each of three consecutive days. The observers had to identify whether a conflict occurred in each piece of film, and if it did, to make a drawing of it and allocate it a grade between 0 and 4. The definitions of these grades were based on Older and Spicer (1976).

Inter-observer variability on the third day was measured by a correlation coefficient of 0.68. Intra-observer reliability between day 2 and 3 was measured at 0.75. There were large differences among the observers, in their consistency and detection rate.

Concordance rates on day 3 between observers and the criterion set by expert judgement was 0.67. The majority of the observers agreed with the criterion in 66.7% (48) of the 72 situations.

There were no significant differences between male and female observers, or between drivers and non-drivers.

The Theory of Signal Detection was applied to the data obtained, and the benefit of selectively eliminating various percentages of observers was examined.

Introduction

Our objectives in this study were to see whether the identification, classification and recording of traffic conflicts could be adequately carried out by the type of personnel typically employed by local authorities as temporary traffic enumerators. If sufficient ability existed in these people, then we could advocate that these studies in future be carried out by such part time personnel. This would release full time investigators from this time-consuming work and would probably encourage more studies to be undertaken. From the relatively simplistic pool of enumerators which many authorities use, a number could be turned into a useful, objective and sophisticated team capable of carrying out all manner of skilled observation techniques including conflict studies.

The specific issues to be resolved were:

a) what is the extent of the variability between observers in the detection and accuracy of grading conflicts?

b) how consistent are observers as regards the detection rate and the accuracy of their gradings?

c) is there evidence to suggest that conflicts at some locations are more difficult to identify than at others?

d) are there any sex differences in the ability to detect and grade accurately?

e) do drivers have any more or less difficulty identifying and classifying conflicts than non-drivers?

Method

Observers

A total of 42 student observers were recruited from within the University. No selection procedures were applied as the whole nature of reliability and variability among observers was under study.

There were 4 separate training sessions with between 7 and 14 people in each. The size of the screens displaying the films limited the maximum size of the groups.

Instructions to observers

After each observer had read the written instructions, the experimenter went through them again verbally in greater detail. They explained what a conflict was and briefly the reasons for studying conflicts

as an aid to accident investigations. The grading system (based on Older and Spicer 1976) was explained and diagramatic examples presented.

A demonstration film was shown to give an example of a conflict in real life and was presented several times with opportunity for discussion and queries. The speed could be reduced and the film stopped on request.

One minute of film of each of the locations was shown and the layouts explained. After each, a number of trial pieces of film of similar length to those used in the experiment proper were shown, and the observers followed the events from pre-prepared example sheets with diagrams and a written explanation of each conflict together with the appropriate grading.

Finally the trainee observers saw 6 trials (2 of each location) to practice drawing and writing down what they had seen. Each trial was repeated in slow motion with the experimenter pointing out the incident and discussion was allowed.

Before each of the 6 films, the experimenter told the observers which site it would be. There were 12 clips in each film, all from the same site i.e., 2 films were of site 1, 2 of site 2, 2 of site 3. The 6 films were shown in a different random order on each day of the experiment. The observers saw 72 clips each day, a small number of these showing traffic but no conflicts.

Results

Inter-observer variability

The degree of correlation between observers increased during the experiment i.e., observers had a higher level of agreement among themselves on the second day when compared to the first, and was highest on the third day of the experiment (Figure 1)

Day	1	2	3
r	0.61	0.67	0.68

FIGURE 1. Average inter-observer correlation coefficients across days

On the final day (Day 3), the overall correlation between observers was 0.68. Although in statistical terms this figure is fairly good (N = 42), whether this figure is acceptable in general observation work in the field is questionable.

Intra-observer reliability

Figure 2 shows the average correlation coefficients (N = 42) i.e., the degree to which an observer agreed with himself when viewing the same incidents on different days.

Days	1 - 2	1 - 3	2 - 3
r	0.65	0.67	0.75

FIGURE 2. Average intra-observer correlation coefficients across days

The correlations between days 1 and 2 was lowest, but between days 2 and 3 had improved considerably. There were large differences between observers. The highest correlation between days 2 and 3 was 0.91, and the lowest 0.30. Poor quality observers greatly influenced the results and indicate the importance of selection. When observers were ranked for position according to their correlation coefficients and these ranks correlated for Days 1 - 2 and 2 - 3, r = 0.73 i.e., those who scored highest on Day 1 were most likely to score highly on Days 2 and 3, and vice versa.

Comparison with criterion

Each conflict was judged by two "experts" and a criterion grade set for each. Each observer's gradings were compared to this criterion. The correlations showed a trend towards better agreement on the third day than on either of the two previous days.

Days	1	2	3
r	0.60	0.65	0.67

FIGURE 3. Average observer-expert correlation coefficients' across days. (N = 42)

The greatest difficulty for observers appeared to be in distinguishing between grades 1 and 2. 76.5% of those incidents with a criterion grade of 2 were allocated a lower grade by the majority of the observers. Incidents graded by expert judgement as non-conflicts and as grade 1 were correctly graded by 92.3% and 93.3% of the observers respectively.

The problem can be divided into four possible causative factors:

i) category grades 1 and 2 were not mutually exclusive enough i.e., too subjective.

ii) genuine misinterpretation of the incident due in part to the form of the presentation (film) and partly to its short and singular presentation.

iii) a down-grading of the incidents because of the prior knowledge that no accidents occur in the films.

iv) insufficient awareness of the subtlety of the severity of avoidance manoeuvres, possibly due to insufficient instruction and practice in hazard perception.

If the latter could be improved, then it is thought that this in turn will help counter the problems involved in the first three.

Comparison of locations

There was evidence from the results to suggest that conflicts at one of the locations were

a) easier to detect than at the other sites and

b) more reliably graded.

When questioned verbally, most of the observers agreed that incidents at this particular site were easier to identify and record.

Sex differences

No significant differences were found between the results of male observers when compared to female observers.

Driver/non driver differences

There were no significant differences between drivers and non drivers in the grading of the incidents. As the observers were all under 27 years of age, and most were between 18 and 21, the amount of experience among the driver group would be quite small. Most did not own their own car and therefore did not drive regularly. This may account for the result. It is possible that there might be a difference when older observers are considered, where the drivers are regular motorists with some years experience.

Observer selection

As poorer quality observers greatly influenced the overall results of the group, we examined the effect of eliminating various percentages from the data.

Figures 4 and 5 show the results of

i) looking at performance on Day 3 and eliminating various percentages of observers

ii) looking at performance on Day 1 and then eliminating the poorer percentages of observers from Day 3 according to their results on Day 1.

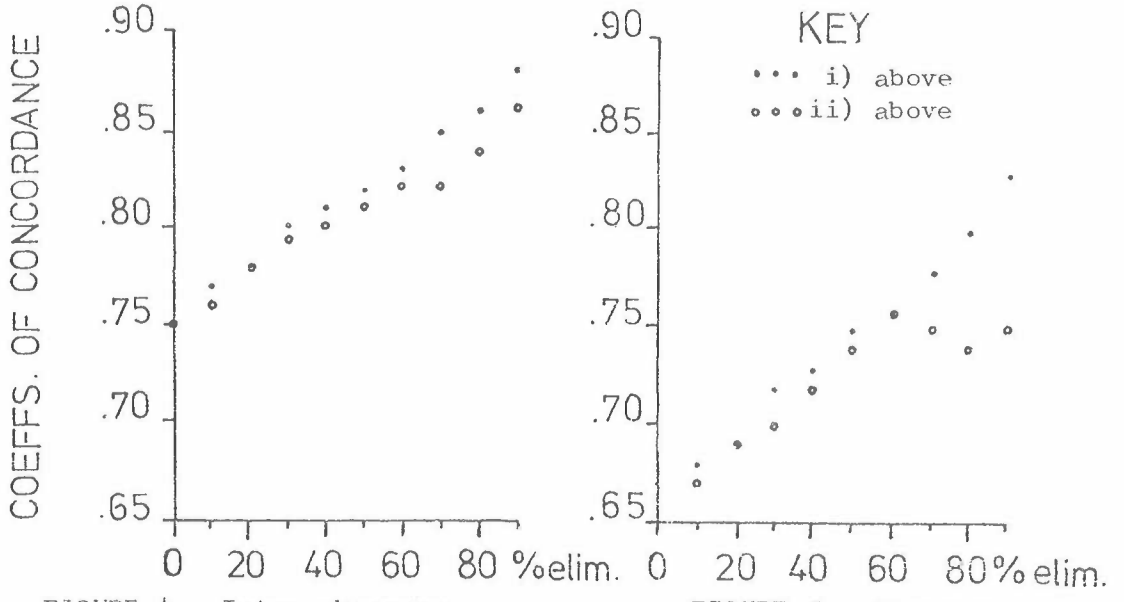

FIGURE 4. Intra-observer correlation coefficients

FIGURE 5. Observer-expert correlation coefficients

This would indicate that it is better to take the trainees through the entire training programme rather than eliminate right at the beginning. Some observers take longer to assimilate the ideas involved in the technique. Nevertheless quite effective selection can be made on the basis of one test after training.

The signal detection theory approach

The Theory of Signal Detection was originally developed to describe the reception characteristics of communication systems, particularly with respect to radar. Tanner and Swets (1954) suggested that it might also be relevant to the detection of signals by human observers in a variety of perceptual tasks. Basically the theory relates the detectability of the signal by an observer to the physical characteristics of the stimulus. A signal is either present or not present and the observer makes a response indicating, according to his judgement, whether the signal is present. Newsome (1974) examined the Theory of Signal Detection concepts of sensitivity and bias as a

possible way of usefully discussing risk taking as a decision process in driving. The detection task for our observers was to correctly identify from the vehicle manoeuvres when corrective or avertive action had been taken. For our purposes

a hit was defined as the correct observation that a conflict had occurred (regardless of grade subsequently given)

a miss was defined as the incorrect grading of a piece of film as not including a conflict when a conflict had occurred according to expert judgement

a false alarm was defined as the identification of a conflict in a piece of film when no such incident had occurred according to expert judgement

a correct rejection was defined as the correct identification and subsequent grading of a piece of film that did not show a conflict situation

The overall distributions from this study are shown in Figures 5 and 6

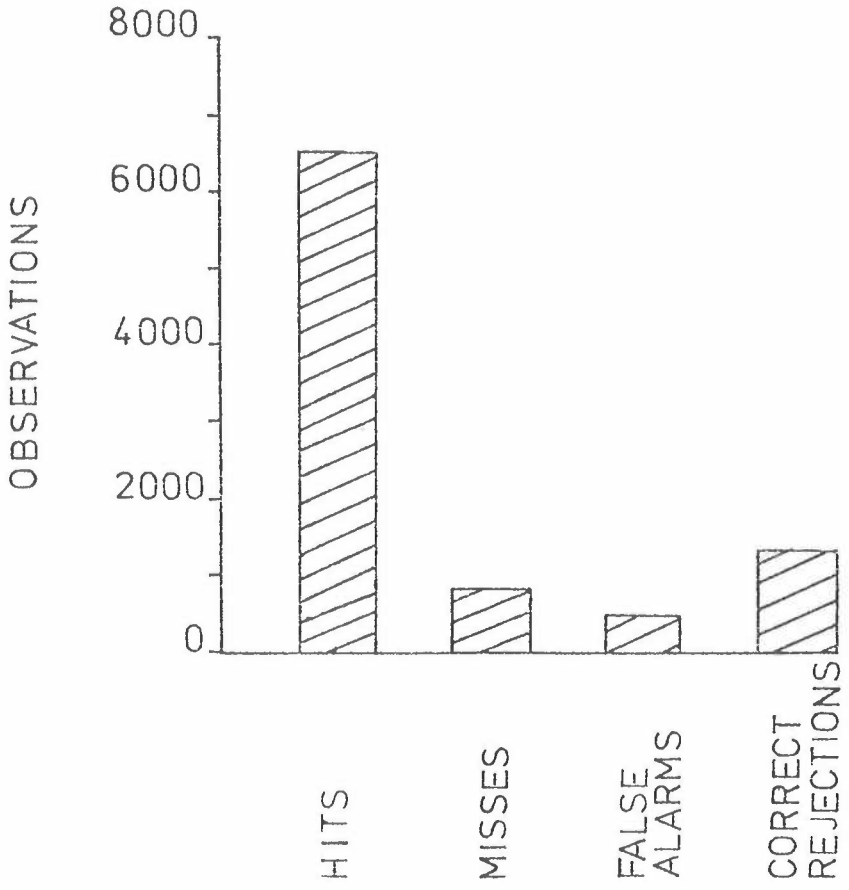

FIGURE 5. Overall distribution and proportions of hits, misses, false alarms and correct rejections.

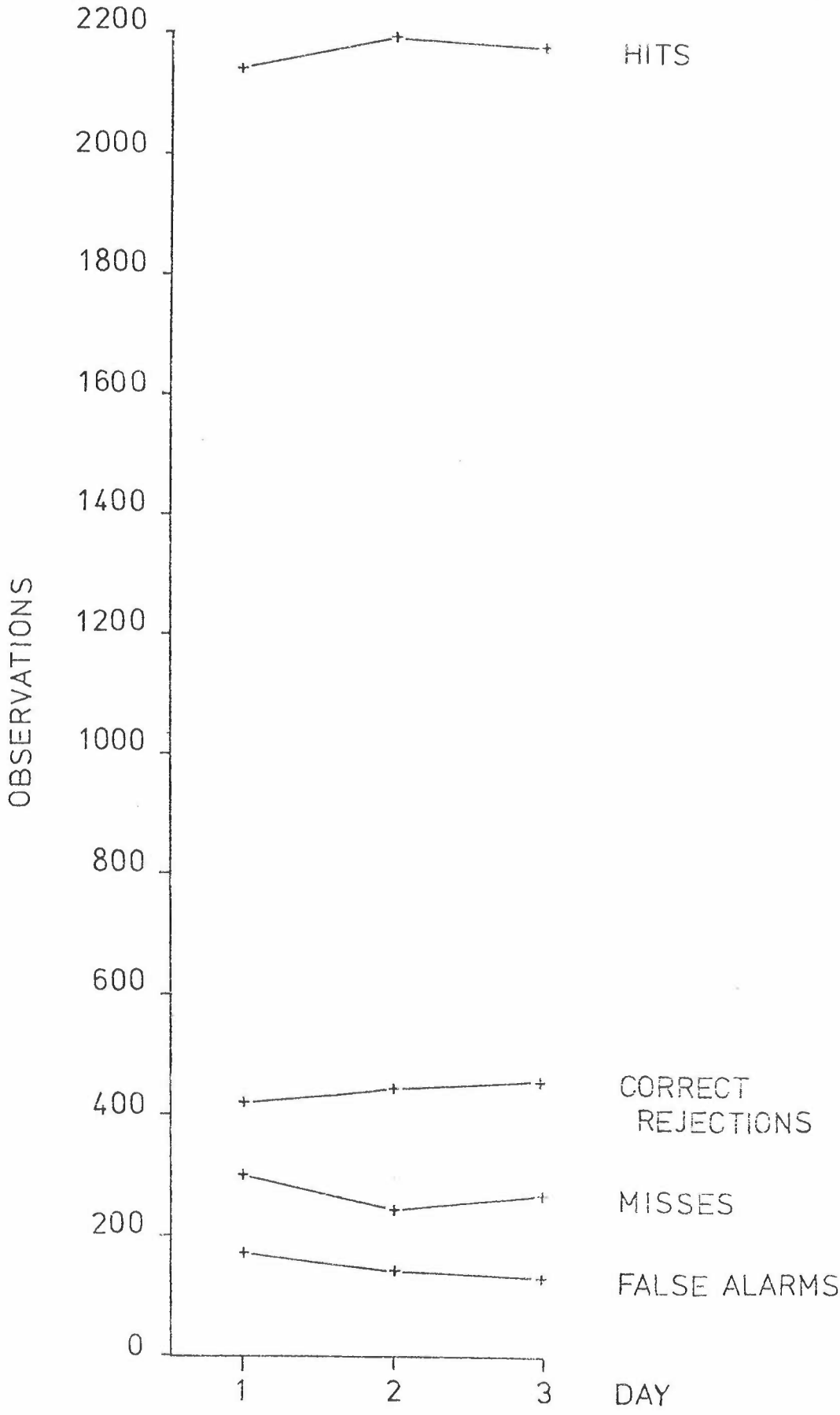

FIGURE 6. Distribution of hits, misses, false alarms and correct rejections across days.

Conclusions

a) With no prior experience, observers can be trained to detect, record and classify those situations known as traffic conflicts.

b) Observers differ considerably in their ability to identify, record and classify traffic conflicts from film, but improved over time and with practice to a maximum of 0.68 (N = 42).

c) Observers are reasonably consistent with themselves. Viewing of the same situations on subsequent days yielded a correlation coefficient of 0.75 (N = 42). This could be improved by observer selection.

d) There is evidence to suggest that conflicts at some locations are easier to detect and grade accurately than at others.

e) No sex differences were observed in the ability to detect and grade conflicts.

f) The question of whether drivers have any more or less difficulty identifying and classifying conflicts than non drivers remain unanswered.

Implications

It would seem that it is feasible to use part-time personnel to carry out conflict studies. This project has demonstrated the levels achieved given a brief (one hour) introduction and training period.

It has provided much valuable information and has helped pinpoint areas which still need to be investigated. It has contributed to our ultimate aim, which is to produce a training manual and associated visual aids for use by local authority traffic and road safety departments to train personnel to carry out traffic conflict observation studies. The results obtained would help in identifying problems precisely and should ensure that remedial measures are more soundly and economically based.

References

Newsome, L. R. Risk taking as a decision process in driving. Transport and Road Research Laboratory, 1974, SR 81 UC

Older, S. J. and B. R. Spicer. Traffic Conflicts - a development in accident research. Human Factors, 1976, 18 (4), 334-350

Tanner, A. P. and J. A. Swets. A decision - making theory of visual detection. Psychological Review, 1954, 61, 401-409