The Reliability of Observed Reintegrative Shaming, Shame, Defiance and Other Key Concepts in Diversionary Conferences.

Nathan Harris and Jamie B. Burton

Table of Contents

Acknowledgments

The authors would like to acknowledge the financial support provided by the Criminological Research Council and the National Drug Crime Prevention Fund. We would also like to thank Yvonne Pittelkow for her statistical advice, John Braithwaite for his support and Charles Barton and the Australian Federal Police for their assistance in the collection of this data.

Introduction

The Reintegrative Shaming Experiment (Sherman, Braithwaite and Strang 1994), which began in the Australian Capital Territory in July 1995, is a comparison between the court system and an alternative to traditional criminal justice interventions, called Diversionary Conferencing. As such, the experiment is primarily focused upon testing, through a randomised trial, differences in the level of recidivism resulting from the two interventions. However, in addition to this goal the experiment also measures a large number of factors which have been predicted by a variety of theoretical perspectives to impact upon criminal activity and recidivism. These factors have been measured using a number of different methods (offender interviews, victim interviews and observation) of which observation has been traditionally considered one of the most difficult to do well. Despite this, observation is seen as a particularly important component of the experiment because of its ability to provide more direct information about what occurs in each of the cases. Therefore, in order to determine how well these concepts could be measured through observation, a study was conducted which tested the inter-rater reliability of the instruments. This report presents the results of that study.

A number of theoretical perspectives were of interest. Perhaps the most important of these, because of it's influence in the development of Diversionary Conferencing, was the theory of reintegrative shaming (Braithwaite 1989). While initially forwarded by Braithwaite in 1989, the theory has since been operationalised by Makkai and Braithwaite (1994) as involving four main facets. These contend that an intervention embraces reintegrative shaming when it involves:

1. Disapproval of the offence while sustaining a relationship of respect;
2. Ceremonies to certify deviance terminated by ceremonies to decertify deviance;
3. Disapproval of the evil of the deed without labelling the person as evil; and
4. Not allowing deviance to become a master status trait

(Makkai and Braithwaite 1994).

In addition to measuring shaming behaviours, and the extent to which they were reintegrative or stigmatizing, it was also of interest to measure whether offenders did display the emotion of shame during proceedings. Measurement of this emotion has traditionally been considered difficult (Scheff and Retzinger 1991) and so it was of considerable interest to determine whether observation instruments were able to measure it reliably.

A second concept considered important to understanding the success of both conferencing and court was the offender's level of defiance (Sherman 1993). Of significance was how offenders reacted towards the process, for example whether they reacted angrily, outwardly challenging the process, and what factors contributed to this response.

Beyond these central theoretical perspectives, a large number of other concepts were also seen as important to understanding conference and court processes. These include whether the offender was remorseful, whether the consequences of the act were explored meaningfully, how much discussion of a resolution occurred and what it focussed upon, whether the case was procedurally just, and the manner in which the offender was treated.

In an attempt to measure these concepts through observation, two separate approaches were developed. The first was a Global Ratings Questionnaire which was completed by the observer at the end of each case. The questionnaire consisted of eight-point scales which were organised in sections relating to each of the concepts outlined above. These questions were answered by the observer immediately after the case based upon their general impression of events.

The second instrument used was the Systematic Observation Instrument. This instrument consisted of eight categories: 'respect for the offender', 'disapproval of act', 'disapproval of offender', 'offender apologises', 'offender is forgiven', 'offender is defiant', 'consequences of act', and 'outcome'. Each page of the instrument consisted of a matrix with the eight categories listed down the left hand side column and 15 response columns to their right (see Appendix 1). Once a relevant theme was first detected, a record was made in the row next to the appropriate category. For example, if someone said something respectful of the offender (eg "He is a really nice kid") then the observer would circle a box in the row 'respect for the offender'. Each observation was put in the next column so that the sequence of observation was recorded. The other important feature of the instrument was that a new observation was only recorded when it differed from the preceding one. This meant that the instrument did not record the intensity, duration, or quality of interactions, only that communications of a certain category occurred in a particular sequence.

The Systematic Observation Instrument was designed to capture a more precise record of significant events and communications during cases. This enabled the instrument to provide information regarding the relative amounts of each type of communication in cases, and also the sequence in which different types of communications occurred (to determine if defiance always follows disapproval of offender, for example).

Both of the instruments were developed during a pre-testing period. As part of this process definitions of the questions in the Global Ratings Questionnaire and the categories in the Systematic Observation Instrument were developed and more tightly defined. A codebook for the Systematic Observation Instrument was written that specified how the instrument was to be used and what should be included and excluded from each of the categories. This process involved observers discussing and reaching agreement on how the different questions should be interpreted so as to maximise the reliability of the instruments and to develop a consistent training procedure for new observers.

The purpose of this study is to determine the inter-rater agreement of the instruments. This will achieve two things. The first is to identify the most reliable measures of each concept, which is part of the process of further refining the instruments by identifying areas that need further work. The second aspect of this study will be to provide data on the reliability of those items which are finally used in the Reintegrative Shaming Experiment.

Method

Sample

The sample of cases included 45 observations. These were made up of 15 property, violence or drink drive court cases, 15 property or violence conferences and 15 drink drive conferences. These categories largely correspond to those observed in the Reintegrative Shaming Experiment (RISE). The only difference being that the reliability study included some adult property cases whereas RISE includes only juvenile property cases (see Table 1).

The basic sampling strategy was to cover all the conferences that fell within the relevant categories during the reliability study period. This, however, proved to be unworkable due to unreliable notification of conferences which are organised independently by facilitators in each of Canberra's four districts; without any central information point conference notification was poor. Consequently the basic sampling strategy was adapted to suit these conditions. All the conferences that both observers could get to were included in the reliability study. The primary reason that conferences were not attended was that the observers were not notified of the conference, however a few were missed because one or both observers were unable to attend. All the observations occurred between April 12 and September 14, 1995.

Table 1: Types of cases observed in the reliability study.

	type of case	number observed	no. involving adults
Conference	theft / shopstealing	10	2
	criminal damage	4	0
	assault	1	0
	drink driving	15	15
Court	drink driving	8	8
	theft / shopstealing	5	1
	assault	2	1

The court cases were attended over the same period. Court was attended on days where both observers were available and there were relevant cases. On days where court was attended, the first two cases in any of the relevant categories (property offences, nondomestic and non-sexual violence offences or drink drive offences) were included in the reliability study. On several occasions only one cases was observed because it was the only relevant case on that day.

Observers

Three observers were used to collect the reliability data. Half way through the conference observations one of the observers changed. This meant 15 conference cases each were observed by a different combination of observers. Observer one attended all 45 cases, observer two attended 30 of the cases (15 court cases and 15 conference cases) and observer three attended 15 of the cases (all conference cases). For convenience observer 1 will be called rater 1 and observers 2 and 3 will be called rater 2. All three observers were male and aged between 20 and 30 years old.

Measures

The reliability of two questionnaires was tested. The first questionnaire was the Systematic Observation Instrument which observers filled out as the case proceeded. The second instrument was the Global Ratings Questionnaire which the observer completed after the case was finished.

Procedure

For each case two observers would attend and complete the two instruments. During the conference or court case the observers would independently fill in the Systematic Observation Instrument and afterward both observers would complete the Global Ratings Questionnaire. Several steps were taken to make sure that the observations were completed independently. During the case observers sat apart where this was possible and did not communicate about the observations. Similarly, the global observations were completed before the observers discussed the case.

Results and Discussion

Global Ratings Questionnaire

In order to measure inter-rater agreement a number of statistical methods were considered. One was the Pearson Product Moment Correlation coefficient which was used to describe the strength of linear association between raters' scores. Although reported, for a number of reasons correlations are an imperfect measure of agreement and should be treated with caution. While the correlation coefficient takes into account variation between the observers it does not provide information about the slope of the relationship or its origin. The implication of this is that a perfect correlation of 1 can be achieved even though two observers are recording entirely different responses, or indeed responses at different ends of the scale. It is equally possible for a perfect correlation to be obtained when one observer uses a very small range of the scale while the other uses its full range. An example of this is question 6 (see page 10) where the correlation between raters for court cases is .55, which, although not very high, indicates a moderate amount of association. However the regression line suggests that there is very little agreement. The slope of the relationship is almost flat (.28) so that while the raters varied systematically there was considerable difference between the actual scores.

A visual inspection of the scatterplots allows one to see where the correlation coefficient is not adequate and also allows one to gain an impression of inter-rater agreement. However, it was also important to develop a statistic to describe more accurately the extent of agreement. This has been done very simply by calculating the percentage of cases in which the raters scored within 1 of each other.

$$\text{\% of agreement} = \frac{\text{n of cases where the absolute value of (rater 1 - rater 2)} < 2}{\text{total n of cases}}$$

Thus agreement is defined as when the raters give scores that are at the most one different (eg 1 & 1, 5 & 5, or 2 & 3, 7 & 8). This criterion seems appropriate given that perfect agreement would be both unrealistic and so stringent that it would not provide much information about the questions. Equally a greater range of 2 seemed too inaccurate, allowing a difference of 1 and 3 or 5 and 7, for example, to be regarded as agreement. Allowing a difference of one is both a realistic definition of agreement and is stringent enough to allow differentiation between questions. Inspection of the scatterplots suggest that where the agreement score is approximately 70%, agreement between the raters is reasonable.

One limitation of the agreement measure is that it does not take into account the extent of disagreement between raters. Thus it makes no difference to the score if the raters are two or four different.

The cases in the study were observed in a number of different contexts: 15 drink driving (PCA) conferences, 15 property/violence (P/V) conferences and 15 court cases. An initial look at the data suggested that questions were treated quite differently depending upon the context. As a result, the analysis was performed separately for each of the observation contexts. Each context is included in the scatterplots presented below and is represented by a different symbol. A key to the symbols used is included for each plot and the correlation and regression line is presented below in the same order as the key. One limitation of these scatterplots is that if two or more cases occur in the same position only one symbol appears. However, the inclusion of the regression line partially resolves this problem.

A further complicating factor was that the scores for rater 2 consisted of observations by two observers (observer 2 and observer 3). Furthermore an inspection of the data suggested that there were some important differences in the way these two observers rated cases. It was important then to determine how different observer 2 and 3 were from each other before their combined scores could be considered one and compared to rater 1. To do this, scatterplots were produced for all the questions which separated observers 2 and 3. These are included in appendix 3 and will be referred to in the discussion of the questions when there are important differences between the two observers. In comparing observers 2 and 3 correlations scores are reported but must be treated with caution due to the restricted sample sizes.

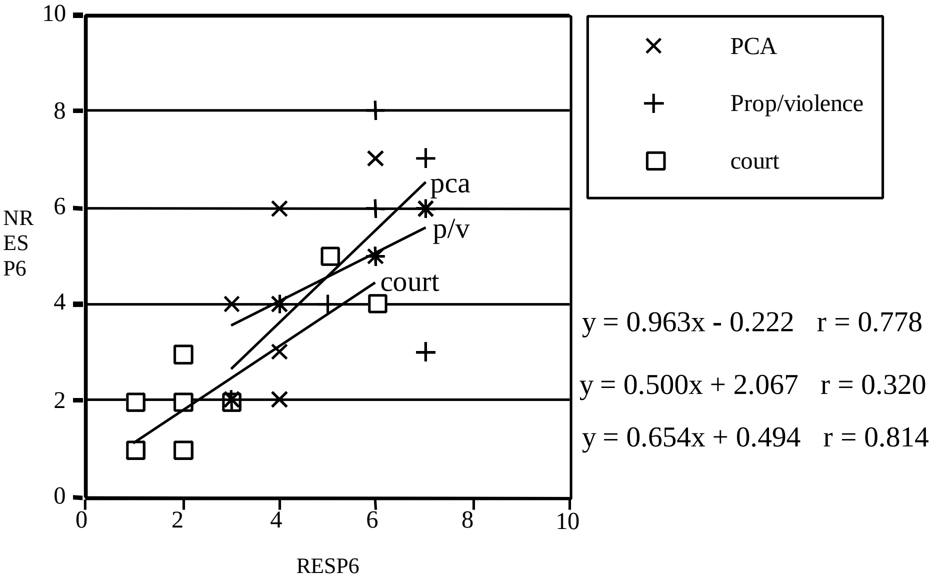
In the following section, all questions, their scatterplots, and discussion of them are included. A summary and discussion of recurring patterns will follow.

Scatterplots of Agreement for the Global Ratings Questionnaire

1. How much reintegrative shame was expressed?

Overall correlation: .79
Agreement. overall: 80% PCA cases: 80% P/V cases
67% Court cases 93%.

Comments: As can be seen below there was high inter-rater agreement in both PCA conferences and court cases but less for property / violence (P/V) conferences. Although agreement is only 67% in the P/V cases this still would appear to be reasonable. There is evidence that observers 2 and 3 differed in P/V conferences (see Appendix 3).

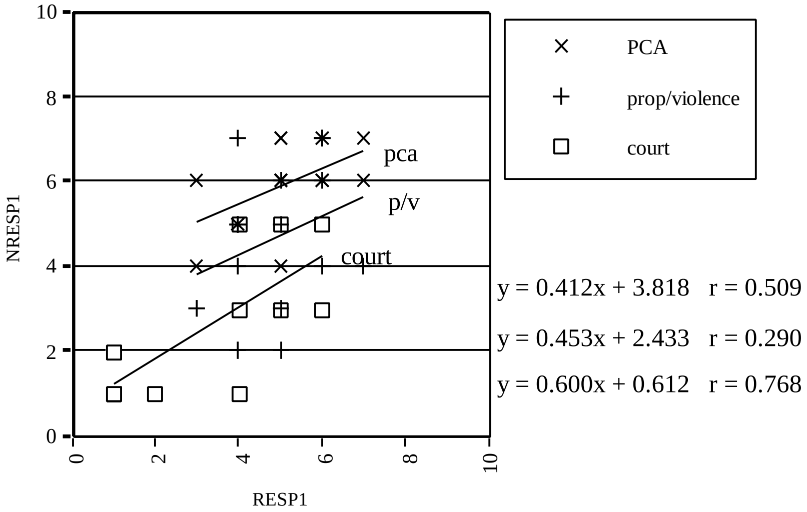


Respect for Offender

2. How much support was the offender given during the conference / court case?

Overall correlation: .71
Agreement. overall: 73% PCA cases: 80% P/V cases 60% Court cases 80%.

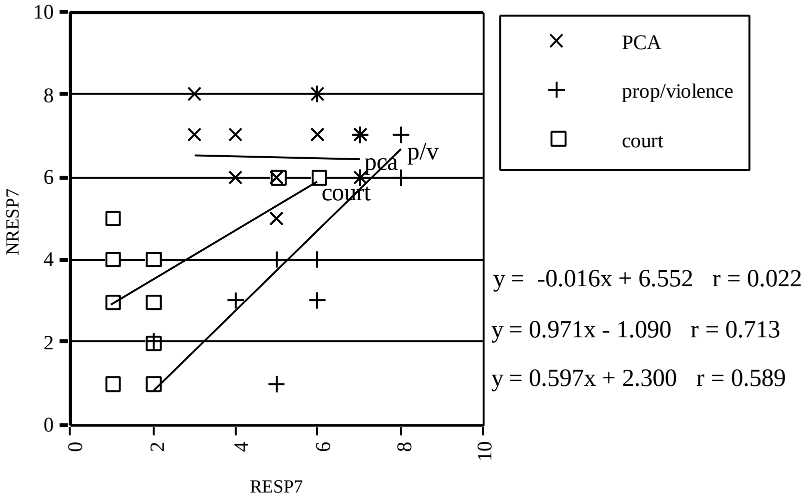
Comments: The agreement for each condition varied between 80% for property/ violence (P/V) conferences and court cases to 60% for P/V cases. The data also suggests that there is a difference between observer 2 and observer 3 (see Appendix 3). Observer 2's scores correlated quite highly with observer 1's in both PCA (.78) and property / violence (.55) conferences, whereas there was a poor relationship between observer 1 and observer 3 in both these conditions (.19 and 0, respectively). It seems that on this question there was a high degree of concordance between observers 1 and 2 but not observers 1 and 3.



3. How reintegrative was the conference / court case?

Overall correlation: .58
Agreement. overall: 56% PCA cases: 67% P/V cases 53% Court cases 47%.

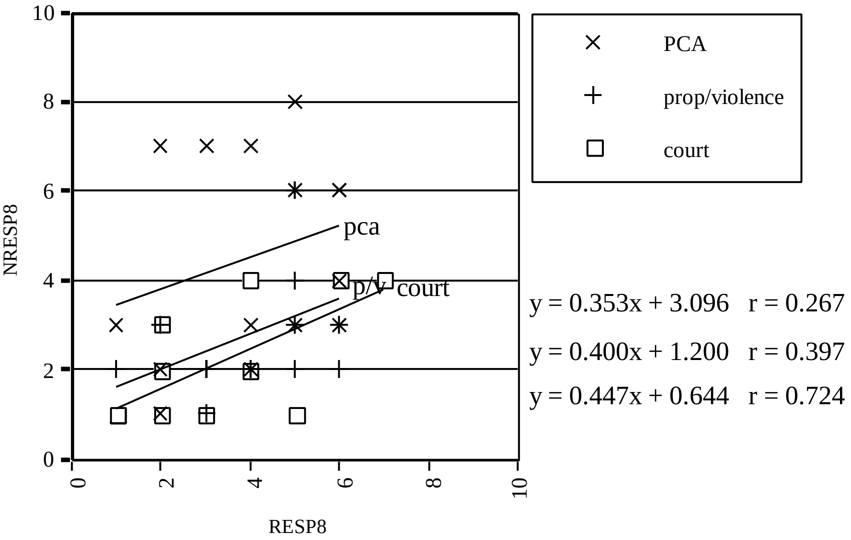
Comments: Inter-rater agreement is fairly low in all observation contexts except PCA cases where it is moderate.



4. How much approval of the offender as a person was expressed?

Overall correlation: .47
Agreement. overall: 53% PCA cases: 40% P/V cases 53% Court cases 67%.

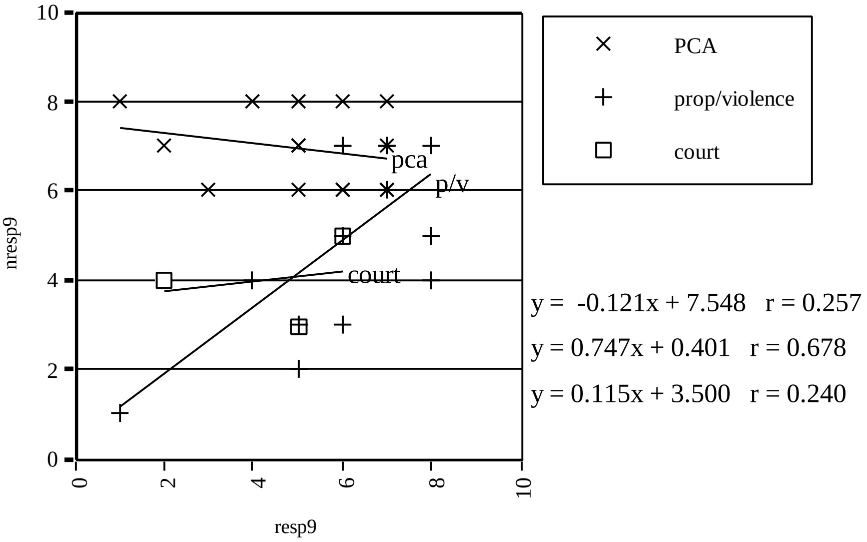
Comments: The scatterplot reveals that there is moderate agreement for court cases, although the regression slope indicates a difference in the use of the scale, but poor agreement in the other contexts. In this case observer 2 and 3 (RESP8) use much more of the scale than rater 1(NRESP8). With the conference cases there was some difference between the observers 2 and 3. In neither case, however, was agreement very high.



5. How much was the offender treated by their supporters as someone they love?

Overall correlation: .19
Agreement. overall: 55% PCA cases: 47% P/V cases 67% Court cases -.

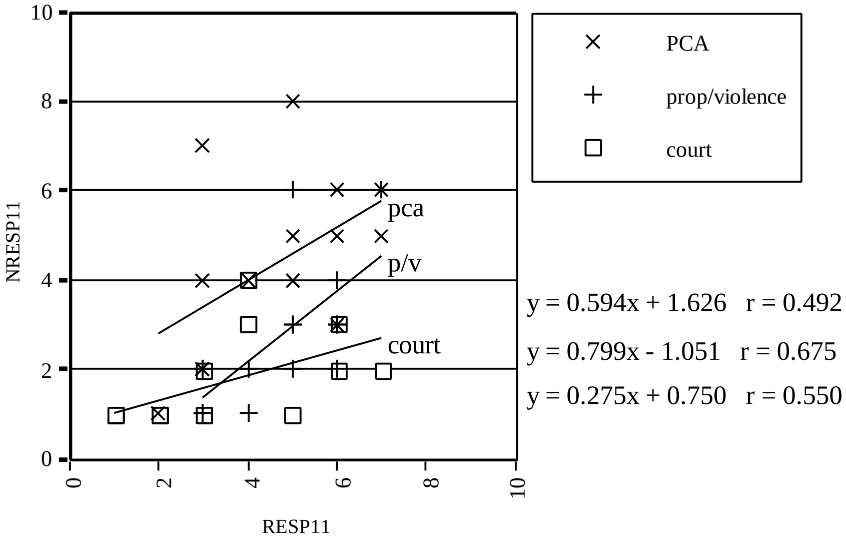
Comments: As the scatterplot shows the agreement between raters is poor for court cases and PCA conferences. The question shows moderate inter-rater agreement in P/V conferences.



6. How much respect for the offender was expressed?

Overall correlation: .56
Agreement. overall: 53% PCA cases: 73% P/V cases
27% Court cases 60%.

Comments: There is reasonable agreement for PCA cases even though there are considerable differences on a number of cases. Agreement is moderate to poor for court cases and very low for P/V cases. Inspection of the relationship (see Appendix 3) reveals a difference in the amount of agreement with observer 2 and observer 3 particularly in relation to PCA conferences. There was high agreement with observer 2 on both V/P conferences (.81) and PCA conferences (.68) but much less so with observer 3 (.52 and .2 respectively).

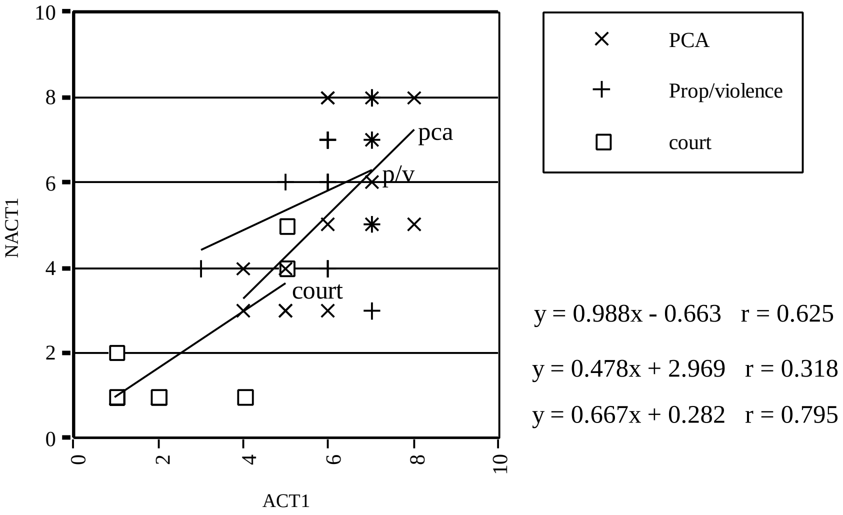


Disapproval of the offender’s act

7. How much disapproval of this <u>type of offence</u> was expressed?

<u>Overall correlation</u>: .83
<u>Agreement</u>. overall: 72% PCA cases: 53% P/V cases 73% Court cases 92%.

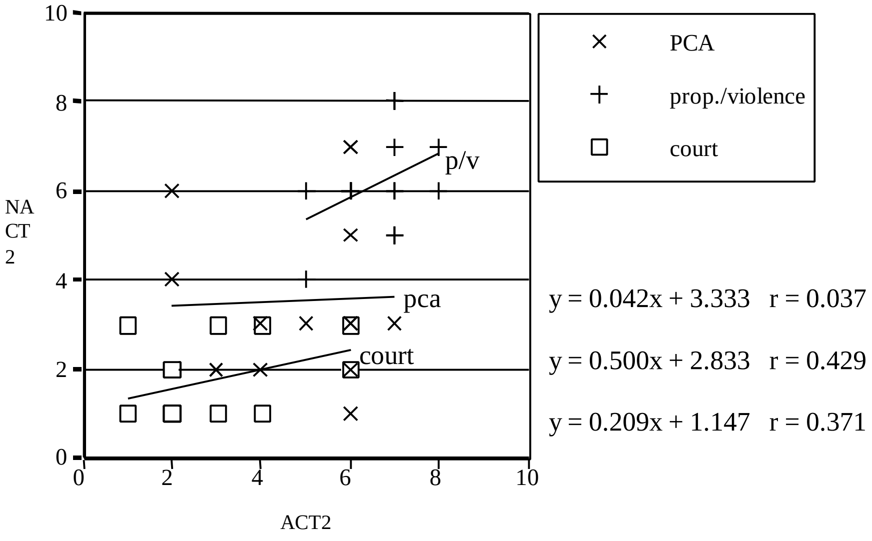
<u>Comments</u>: There is reasonable agreement in P/V conference cases and very high agreement for court cases but low agreement in PCA cases.. For P/V cases there was a considerable difference in the agreement of observer 1 with observers 2 and 3. With observer 2 the inter-rater correlation was .52 compared to 0 for observer 3.



8. How much disapproval of <u>the offender’s act</u> was expressed?

<u>Overall correlation</u>: .67
<u>Agreement</u>. overall: 62% PCA cases: 40% P/V cases 80% Court cases 67%.

<u>Comments</u>: There appears to be little agreement in the way in which the observers have answered this question in PCA cases and only moderate agreement for court cases. There is high agreement for P/V cases.

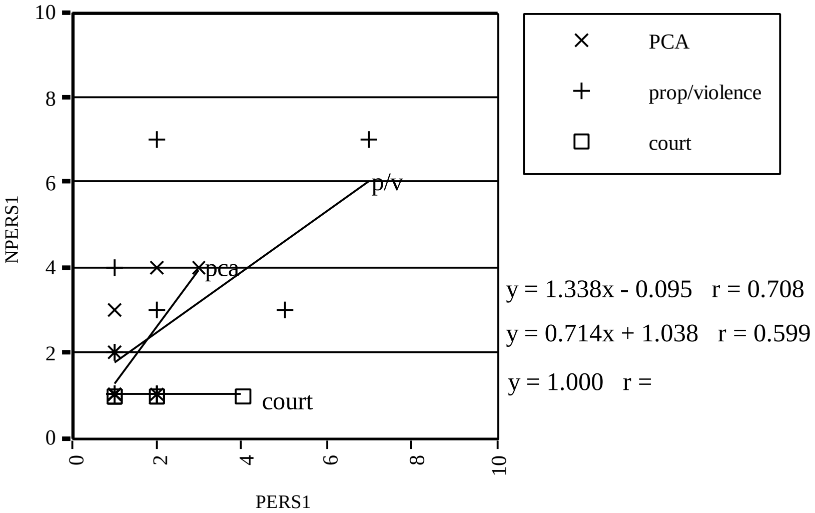


Disapproval of the offender.

9. How much stigmatising shame was expressed?

Overall correlation: .57
Agreement. overall: 87% PCA cases: 87% P/V cases 80% Court cases 93%.

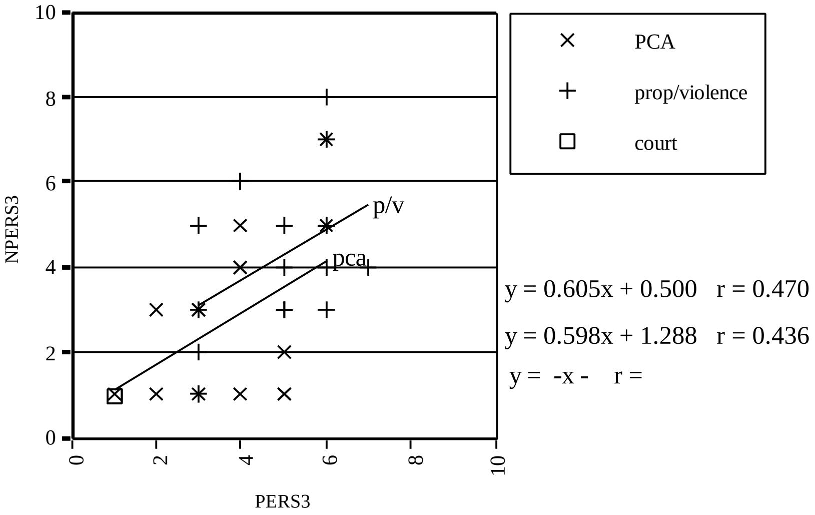
Comments: There is high agreement between observers for all the cases. However this is partially due to the fact that most of the answers indicated that none of this category occurred. Because of the limited range it is difficult to test agreement very well.



10. How much disappointment in the offender was expressed?

Overall correlation: .73
Agreement. overall: 69% PCA cases: 67% P/V cases 40% Court cases 100%.

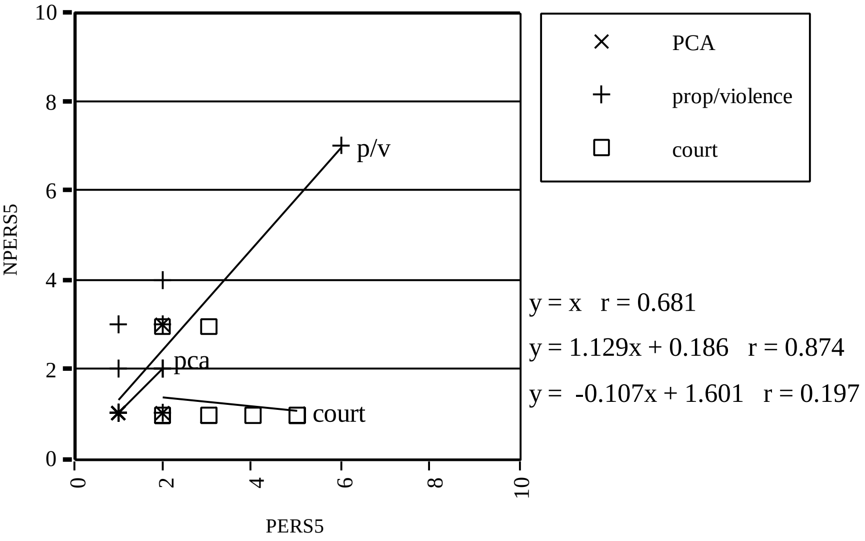
Comments: There is only moderate to low agreement between the observers for the conference cases and not enough variation to assess the inter-rater agreement of the court cases. The moderate agreement for PCA cases was the result of high agreement between observer 1 and observer 2 (.71) and low agreement between observers 1 and 3 (.23).



11. To what extent was the offender treated as a criminal?

Overall correlation: .35
Agreement. overall: 80% PCA cases: 100% P/V cases 87% Court cases 53%.

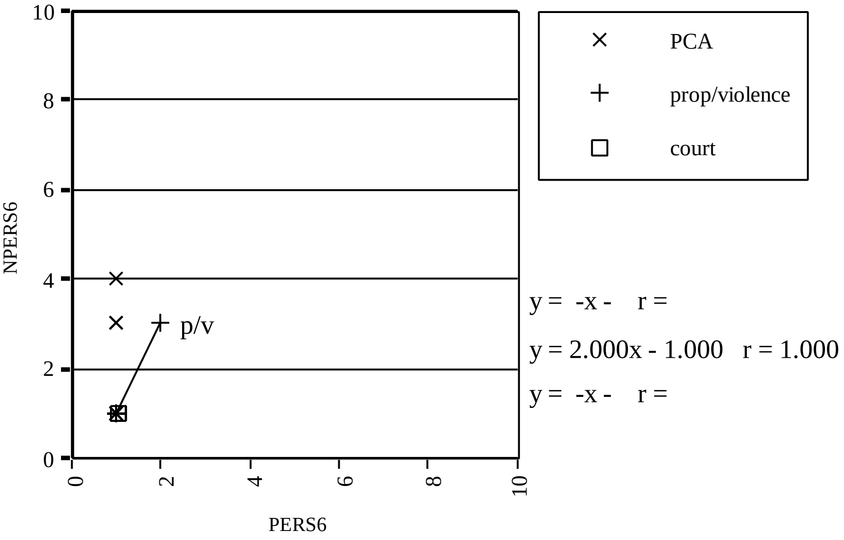
Comments: The scatterplot shows high agreement for conference cases, although the variation is too limited to properly assess agreement. There is low agreement for court cases.



12. How often were stigmatising names and labels (eg, 'criminal', 'punk', 'junkie', or 'bully') used to describe the offender?

Overall correlation: ..38
Agreement. overall: 91% PCA cases: 73% P/V cases 100% Court cases 100%.

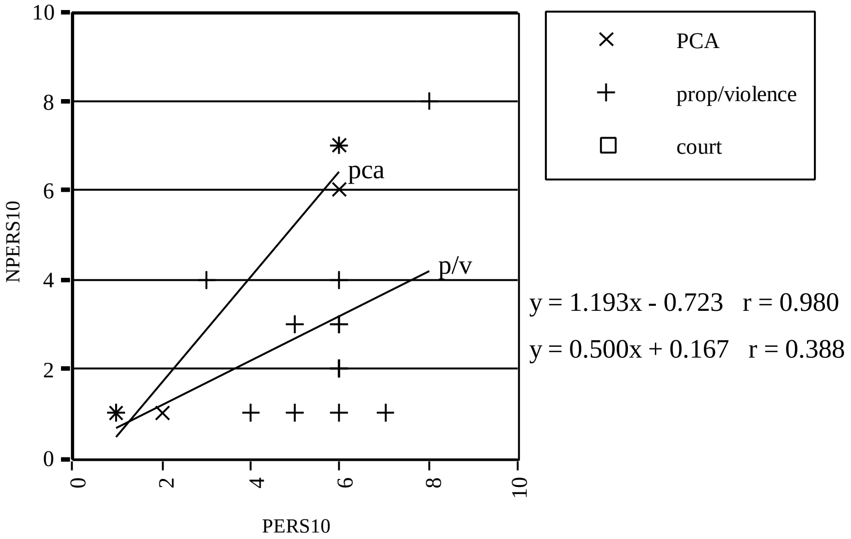
Comments: The scatterplot shows almost no variation on this question. Raters consistently agreed that stigmatising labels were almost never used.



13. How much moral indignation did the victim express about the offender's action?

Overall correlation: .48
Agreement. overall: 42% PCA cases: - P/V cases 27% Court cases -.

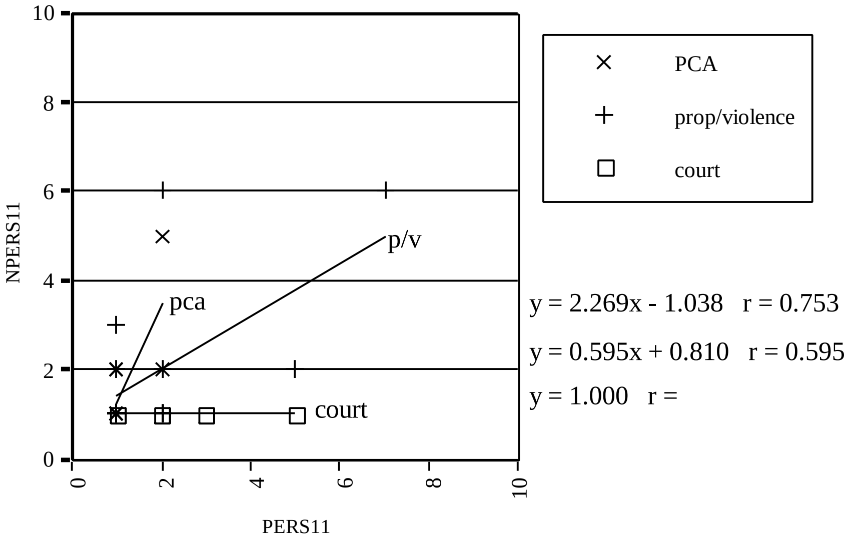
Comments: The item has very poor agreement for P/V conferences and there are too few cases to asses the other contexts. Again there is much greater agreement with observer 2 (.62) than observer 3 (.0) for P/V cases.



14. How much disapproval of the offender as a person was expressed?

Overall correlation: .43
Agreement. overall: 84% PCA cases: 93% P/V cases 80% Court cases 80%.

Comments: In all observation contexts the raters agreed that there is very little of this category but there is not enough variation to adequately test the inter-rater agreement.

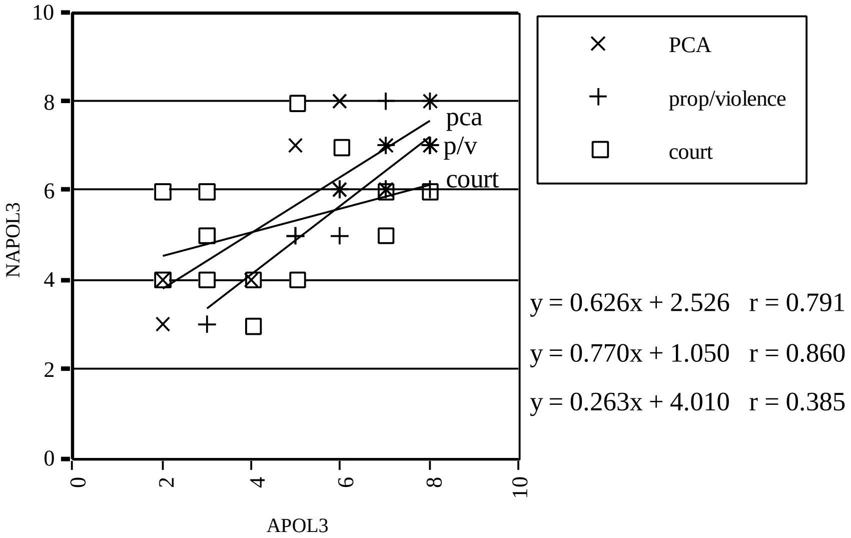


Offender apologises

15. To what extent did the offender accept that they had done wrong?

Overall correlation: .70
Agreement. overall: 69% PCA cases: 73% P/V cases 93% Court cases 40%.

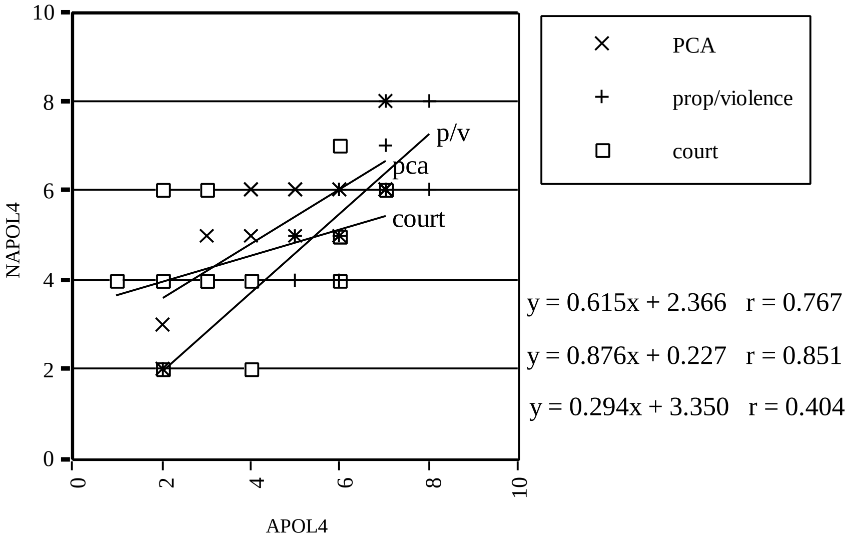
Comments: On this item there is high inter-rater agreement for PCA and very high agreement P/V conferences. This is consistent across both combinations of observers. Agreement for court cases is poor.



16. How sorry/remorseful was the offender for their actions?

Overall correlation: .69
Agreement. overall: 73% PCA cases: 80% P/V cases 87% Court cases 53%.

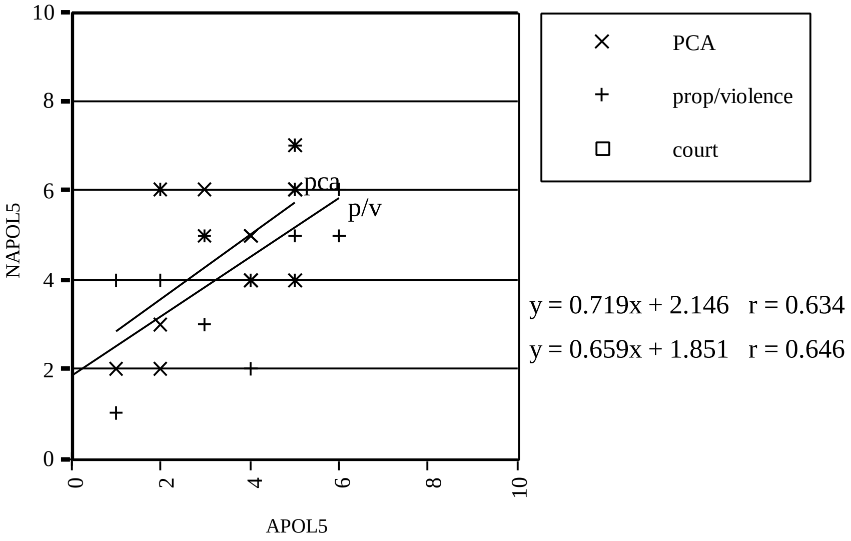
Comments: On this item there is very high inter-rater agreement for PCA and P/V conferences. This is higher with observer 2 than observer 3. Agreement for court cases is poor.



17. When reaching the conference agreement, how severe was the offender on themself?

Overall correlation: .64
Agreement. overall: 66% PCA cases: 73% P/V cases 57% Court cases -.

Comments: There is reasonable agreement for PCA cases. There is a considerable discrepancy between observer 2 (.77) and observer 3 (.17) for the P/V conference cases which have fairly low agreement.

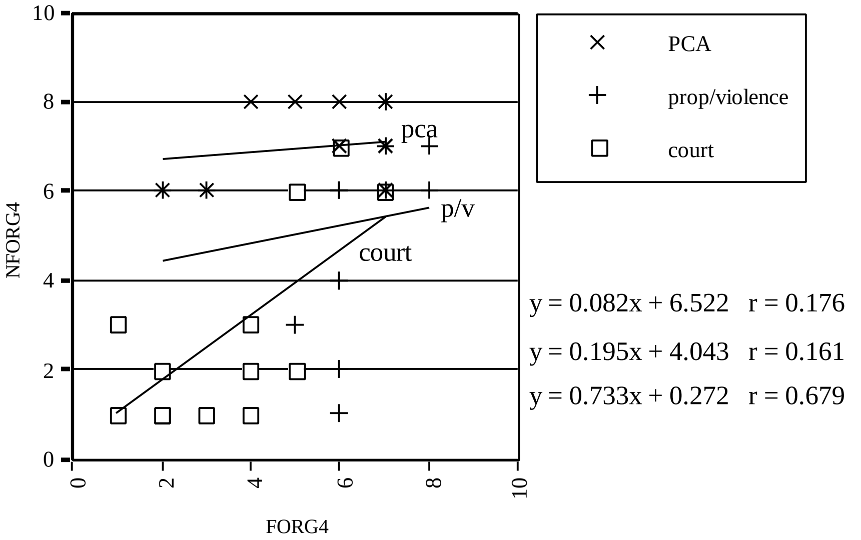


Offender is forgiven

18. To what extent was the offender forgiven for their actions?

Overall correlation: .6
Agreement. overall: 58% PCA cases: 67% P/V cases
47% Court cases 60%.

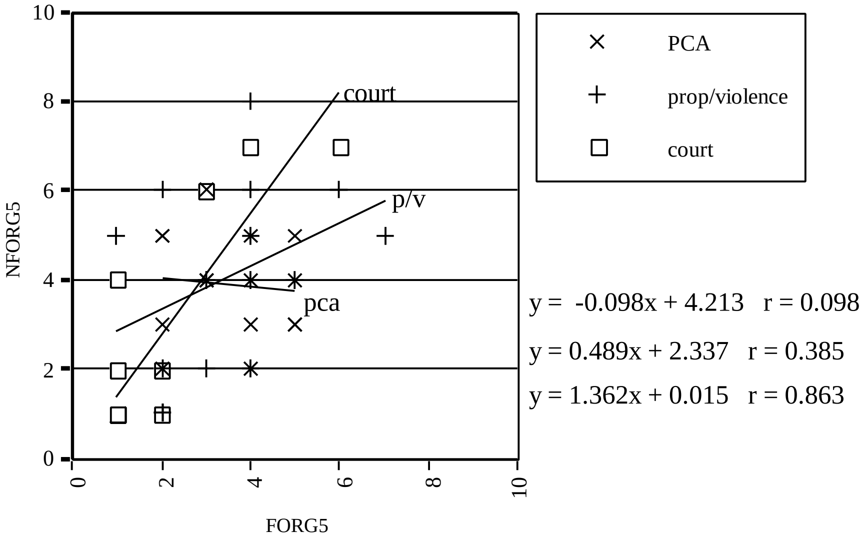
Comments: There is fairly low agreement for all contexts and this is consistent across combinations of observers. The agreement is slightly better for the PCA cases.



19. How clearly was it communicated to the offender that they could put their actions behind them?

Overall correlation: .58
Agreement. overall: 67% PCA cases: 60% P/V cases
60% Court cases 80%.

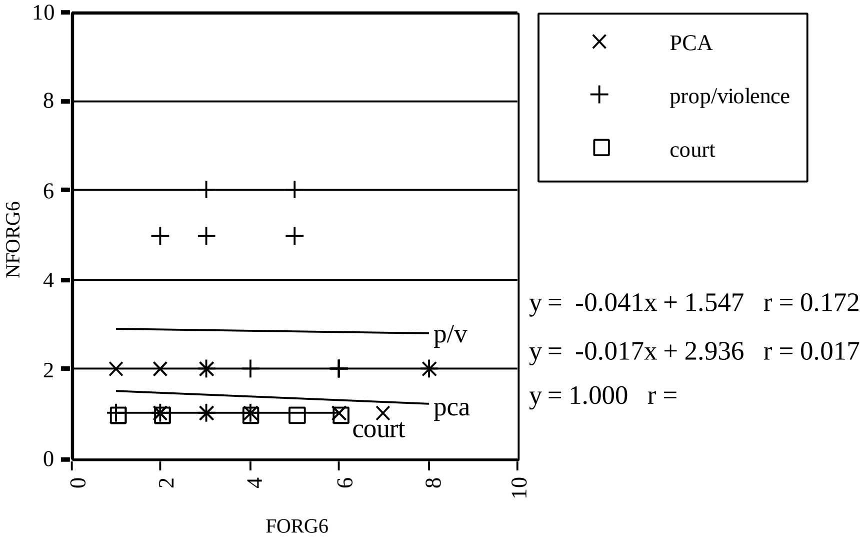
Comments: The plot shows that there is fairly low agreement for conferences cases but good agreement for court cases. The low agreement for P/V conferences is partially the result of differences between the combinations of observers (see appendix 3).



20. How much forgiveness of the offender was expressed?

Overall correlation: .12
Agreement. overall: 51% PCA cases: 53% P/V cases 33% Court cases 67%.

Comments: Across all conditions with both combinations of observers there is little inter-rater agreement on this item. For court cases agreement is marginally better but not much.

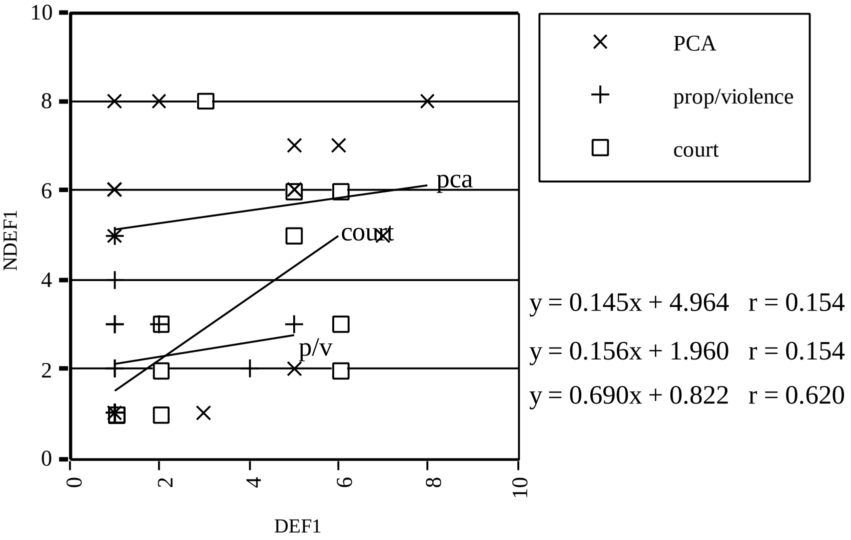


Defiance by the offender

21. How much did the offender claim their actions were accidental or unintentional?

Overall correlation: .44
Agreement. overall: 58% PCA cases: 33% P/V cases 60% Court cases 80%.

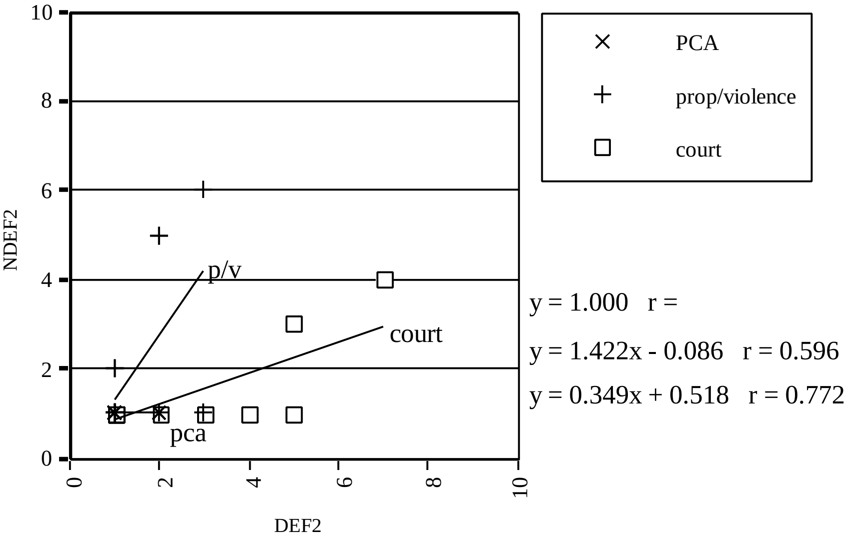
Comments: There is no agreement between raters for conference cases. However there reasonable agreement for court cases.



22. To what extent did the offender hold others responsible for their actions?

Overall correlation: .44
Agreement. overall: 80% PCA cases: 100% P/V cases 73% Court cases 67%.

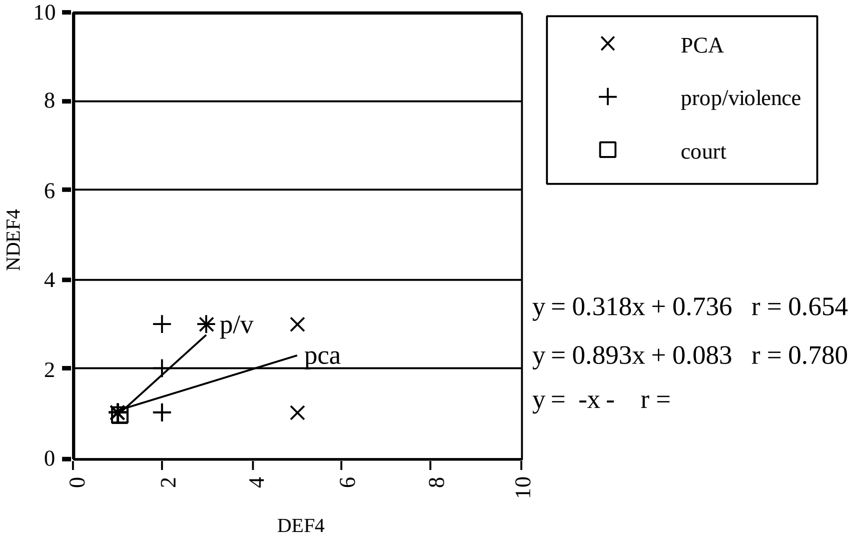
Comments: In all observation contexts the raters agreed that there is very little of this category but there is not enough variation to adequately test the inter-rater agreement.



23. How defiant was the offender?

Overall correlation: .65
Agreement. overall: 96% PCA cases: 87% P/V cases 100% Court cases 100%.

Comments: In all observation contexts the raters agreed that there is very little of this category but there is not enough variation to adequately test the inter-rater agreement.

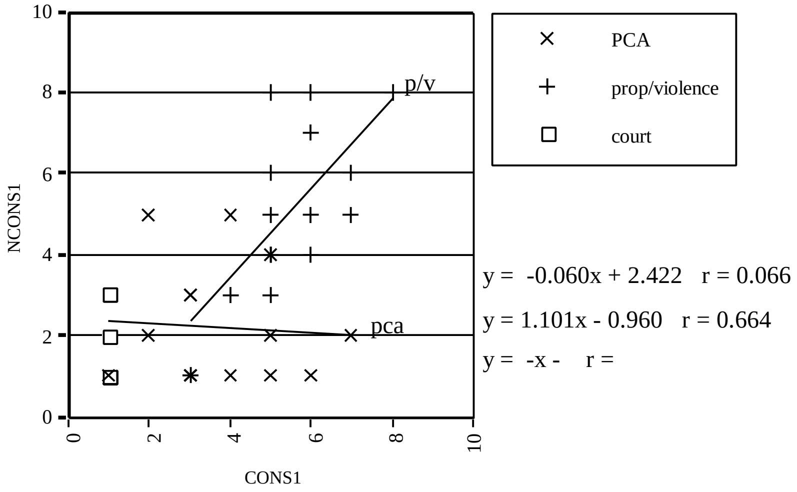


Consequences of the offender's act

24. How emotionally powerful was the account given of the consequences of the offender's act?

Overall correlation: .68
Agreement. overall: 62% PCA cases: 40% P/V cases 60% Court cases 87%.

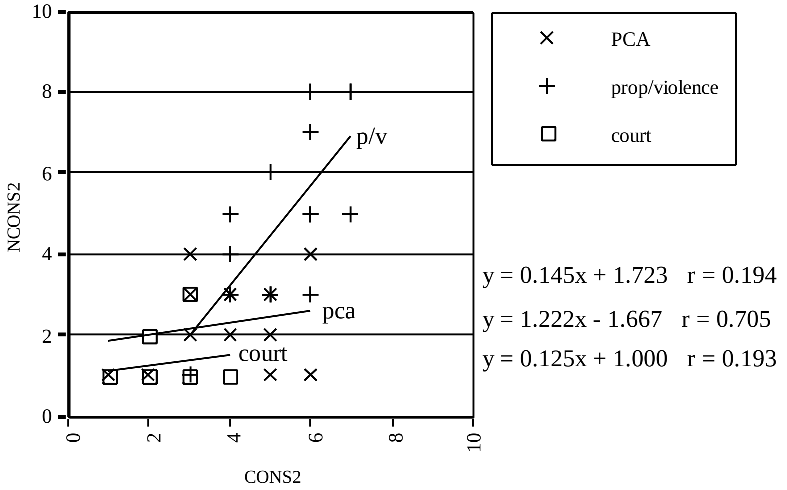
Comments: There is fairly low agreement on this item for P/V cases. This is possibly because of a considerable difference between observer 2 (.74) and observer 3 (.26). There was poor agreement for PCA conferences and not enough variation to assess court cases.



25. How emotionally responsive was the offender to the account given of the consequences of their act?

Overall correlation: .7
Agreement. overall: 60% PCA cases: 47% P/V cases 60% Court cases 73%.

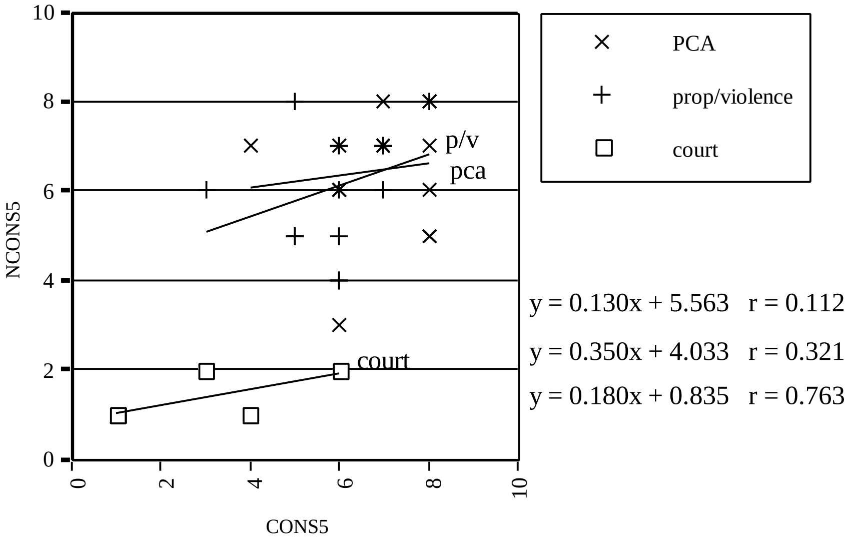
Comments: There is high agreement for P/V conferences and this was consistent across combinations of observers. The agreement for PCA conferences was quite low, though agreement with observer 3 was reasonable (.68) compared with observer 2 (.23). In court cases there was little range or agreement.



26. How much discussion of the consequences (even if not realised) of this type of offence occurred?

Overall correlation: .83
Agreement. overall: 74% PCA cases: 67% P/V cases 73% Court cases 85%.

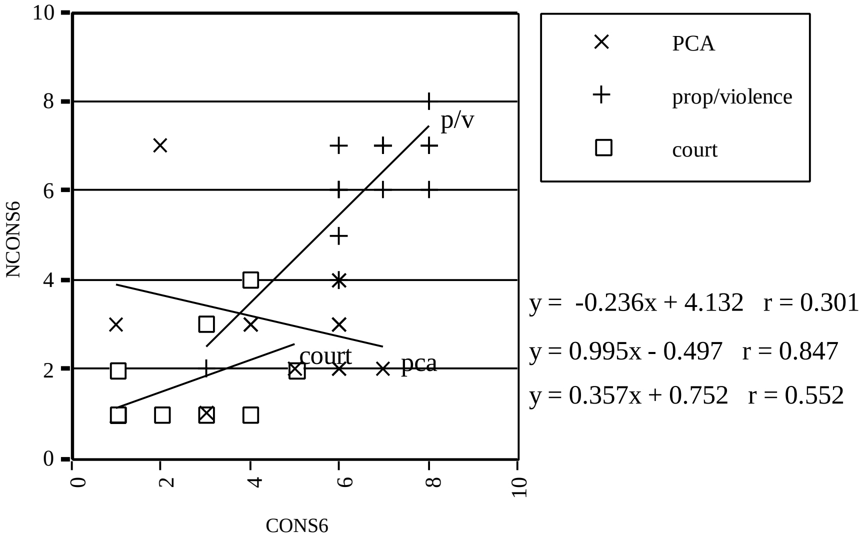
Comments: There is reasonable agreement for P/V cases and moderate agreement for PCA cases. The agreement score for court cases is high but this may be because in most of the cases 'none' of this category occurred.



27. How much discussion of the consequences of the offender's actions occurred?

Overall correlation: .74
Agreement. overall: 62% PCA cases: 20% P/V cases 87% Court cases 80%.

Comments: There is high agreement between raters for P/V conferences and court cases though there is very little agreement for PCA conferences.

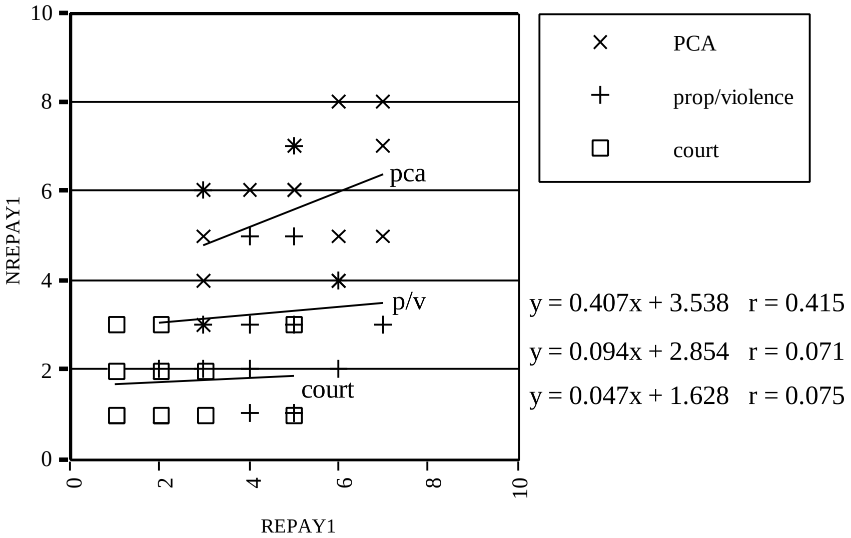


Reparation to victim/Community

28. How much did the offender contribute to the conference / court outcome?

Overall correlation: .56
Agreement. overall: 53% PCA cases: 47% P/V cases 40% Court cases 73%.

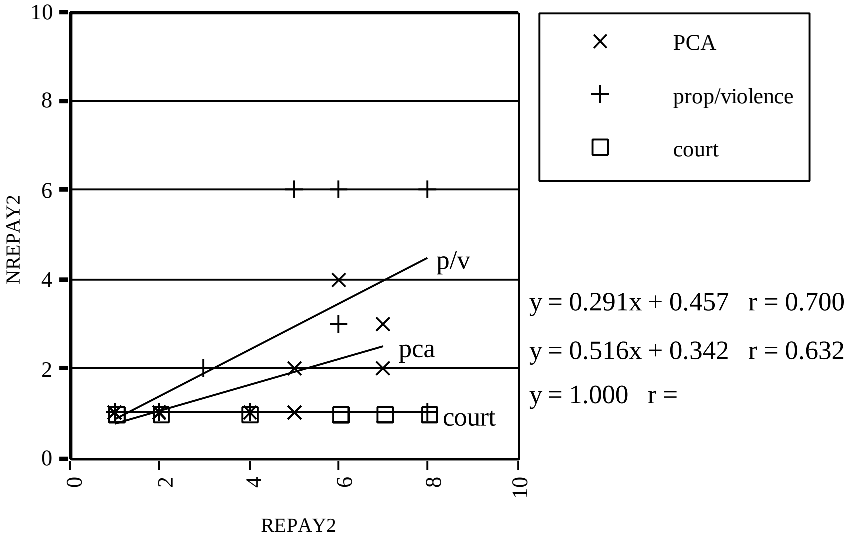
Comments: There is low agreement in conference cases but moderate agreement in court cases. In both the PCA and P/V conferences it is evident that there was much higher agreement with observer 3 (.83 and .45 respectively) than observer 2 (0 and .13).



29. How much was the offender coerced into accepting the conference / court case outcome?

Overall correlation: .36
Agreement. overall: 49% PCA cases: 53% P/V cases 67% Court cases 27%.

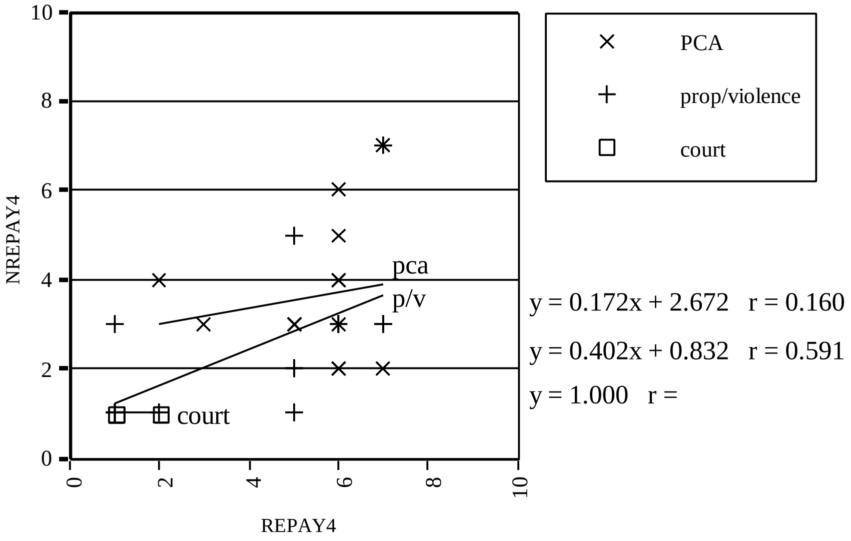
Comments: There appears to be moderate agreement in P/V cases but very little or none in court and PCA cases. On this item there were differences between the combinations of observers with higher agreement between observer 1 and 3 for the P/V cases.



30. How much discussion of paying a debt to the community occurred?

Overall correlation: .69
Agreement. overall: 53% PCA cases: 27% P/V cases 40% Court cases 100%.

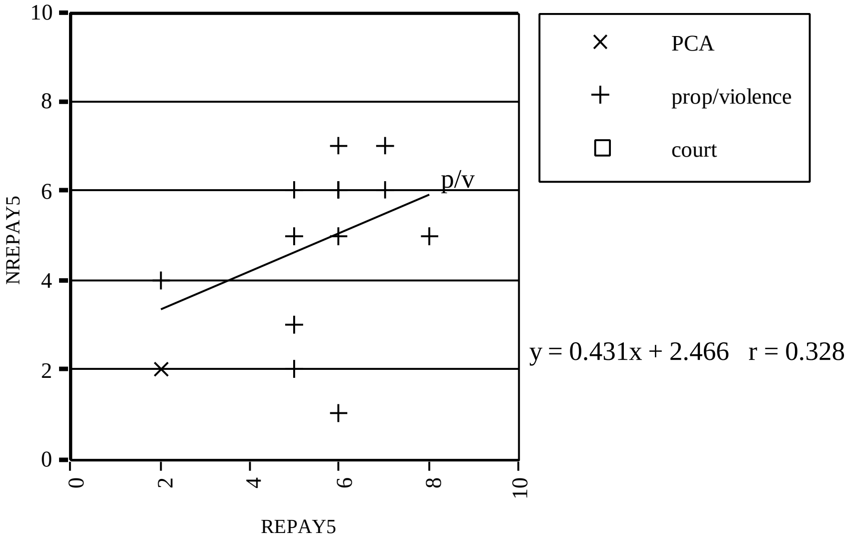
Comments: This question has low inter-rater agreement in both P/V and PCA cases. The range for court cases was too small to assess agreement.



31. How much discussion of reparation to the victim party(s) occurred?

Overall correlation: .33
Agreement. overall: 67% PCA cases: - P/V cases 64% Court cases -.

Comments: This question was only asked in P/V conference cases where there was moderate agreement between raters.

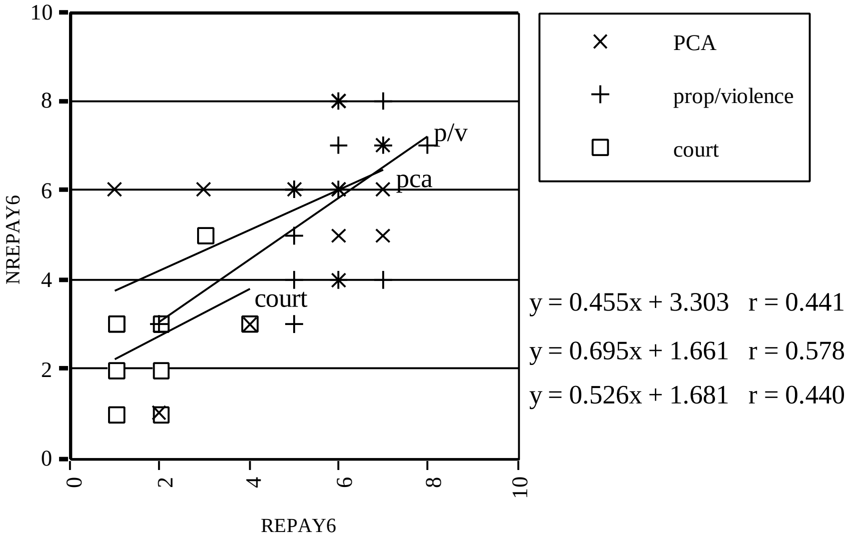


32. Overall how much discussion of reparation occurred?

Overall correlation: .73

Agreement. overall: 67% PCA cases: 53% P/V cases 73% Court cases 77%.

Comments: This question has reasonable inter-rater agreement for P/V and court cases, but agreement in PCA cases is poor.

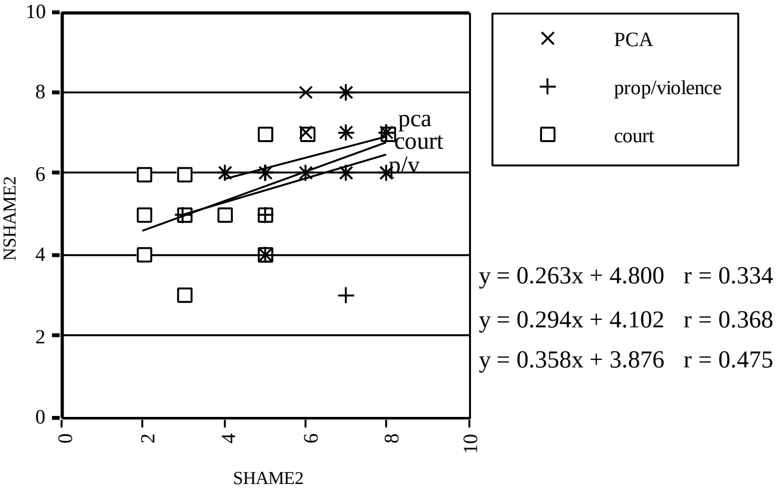


Shame

33. How much responsibility did the offender take for their actions?

Overall correlation: .5
Agreement. overall: 62% PCA cases: 67% P/V cases 73% Court cases 47%.

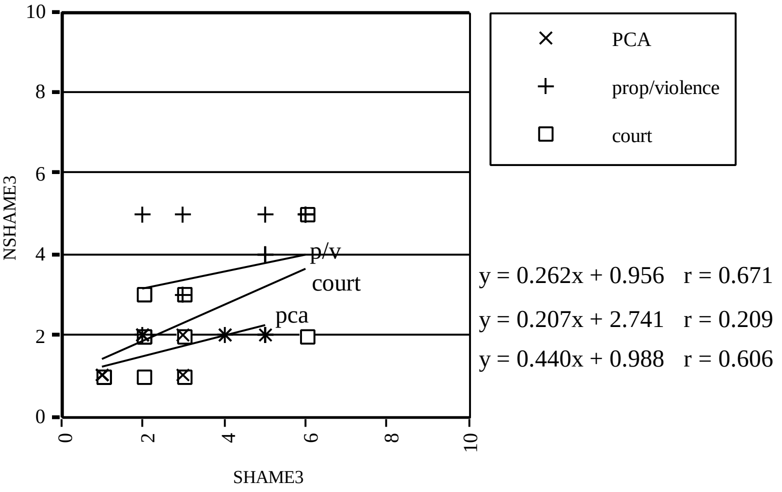
Comments: The inter-rater agreement on this question is reasonable for conference cases but low for court cases. There is some suggestion that over all the cases observers used the scales differently (observer 1 tended to give higher scores but use less of the scale than observers 2 & 3). A closer look at the data suggest that part of the reason for this is that observers 2 and 3 have also used the scale differently. Each agrees reasonably with observer 1 but uses different ranges of the scale, meaning that when combined the relationship is much weaker.



34. How much did the offender retreat from and avoid the attention of others?

Overall correlation: .64
Agreement. overall: 73% PCA cases: 73% P/V cases 60% Court cases 87%.

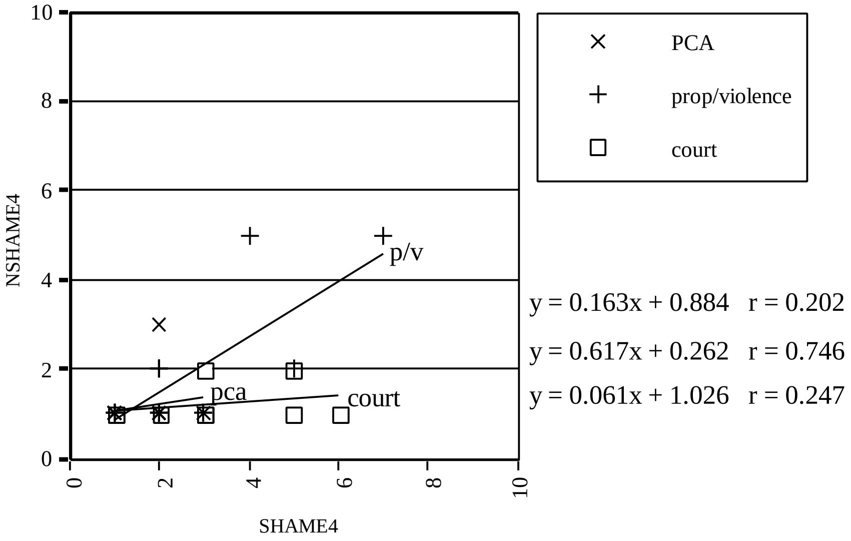
Comments: The scatterplot suggests there is fairly low agreement in the P/V cases but good for court cases. There are considerable differences for the P/V case between observers 2 and 3 with stronger agreement between 1 and 3(.68). PCA cases have a reasonable agreement score but there is limited variation in the sample.



35. How much was the offender's speech affected by irregularities, pauses, or incoherence?

Overall correlation: .53
Agreement. overall: 74% PCA cases: 93% P/V cases 73% Court cases -

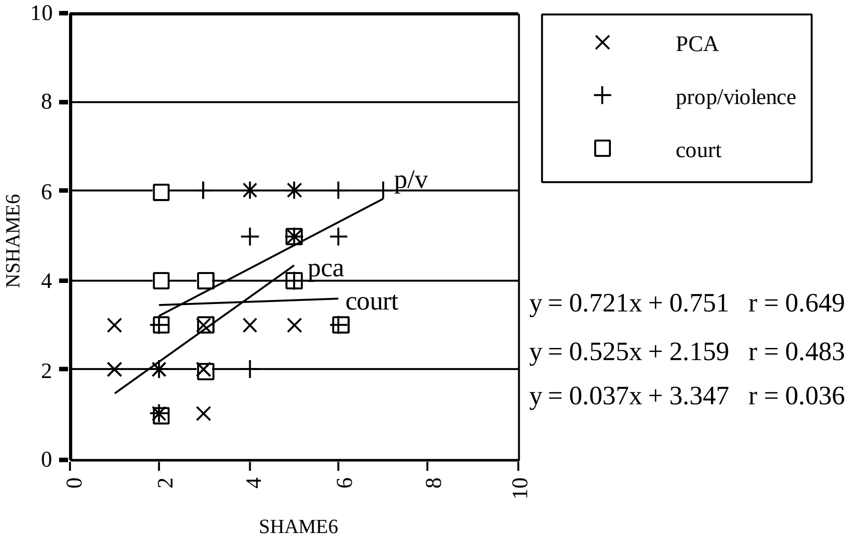
Comments: The conference cases have good agreement scores but this is only because 'none' of this category has occurred. There is to little variation in the conference cases to assess agreement. The court cases have low agreement.



36. How uncomfortable (eg restless, anxious, fidgety) was the offender?

Overall correlation: .48
Agreement. overall: 69% PCA cases: 73% P/V cases 67% Court cases 67%.

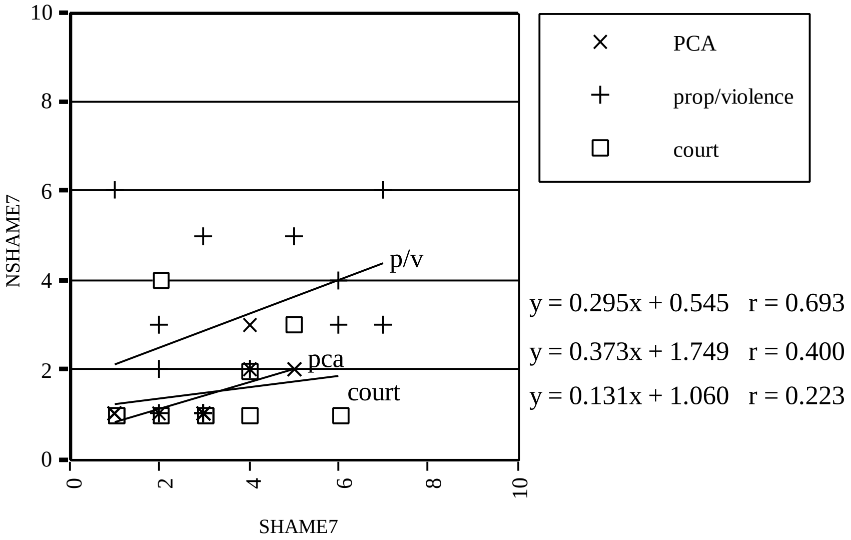
Comments: There is moderate agreement for this item over all types of cases.



37. To what extent did the offender engage in hiding (eg lowering head) and concealing (eg hand covering parts of the face, averting gaze) behaviour?

Overall correlation: .51
Agreement. overall: 47% PCA cases: 47% P/V cases 40% Court cases 53%.

Comments: Agreement on all the cases is low. Agreement for P/V cases may be the result of low agreement between observers 1 and 2. Observers 1 and 3 show quite good agreement for these cases (.83).

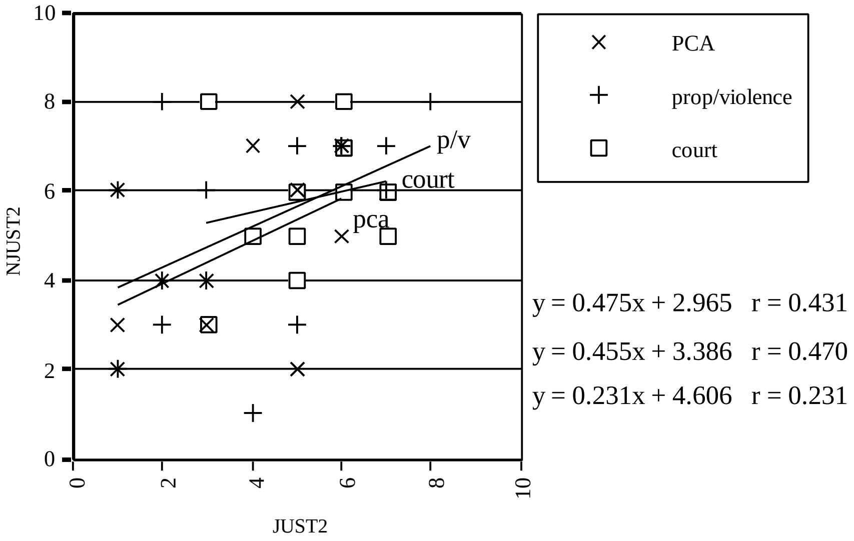


Procedural Justice

38. How much was the offender dominated?

Overall correlation: .46
Agreement. overall: 62% PCA cases: 53% P/V cases
53% Court cases 80%.

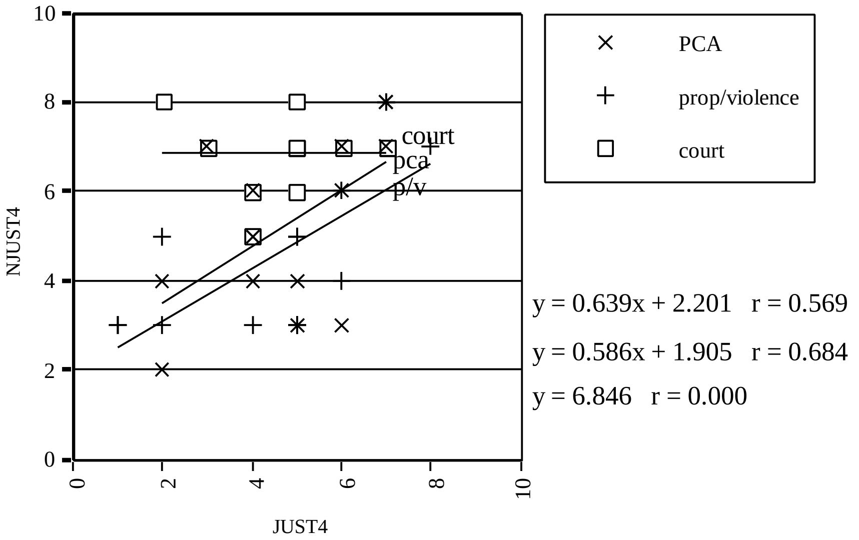
Comments: There is low agreement for the conference cases but high agreement for court cases. Again observers 1 and 3 have fairly good agreement for the P/V cases (.71) but observers 1 and 2 did not.



39. How directive was the facilitator / magistrate?

Overall correlation: .51
Agreement. overall: 56% PCA cases: 67% P/V cases
53% Court cases 46%.

Comments: There is moderate agreement for PCA cases but low agreement for all other cases. For the PCA cases there was higher agreement with observer 3 (.88) than observer 2 (.44).

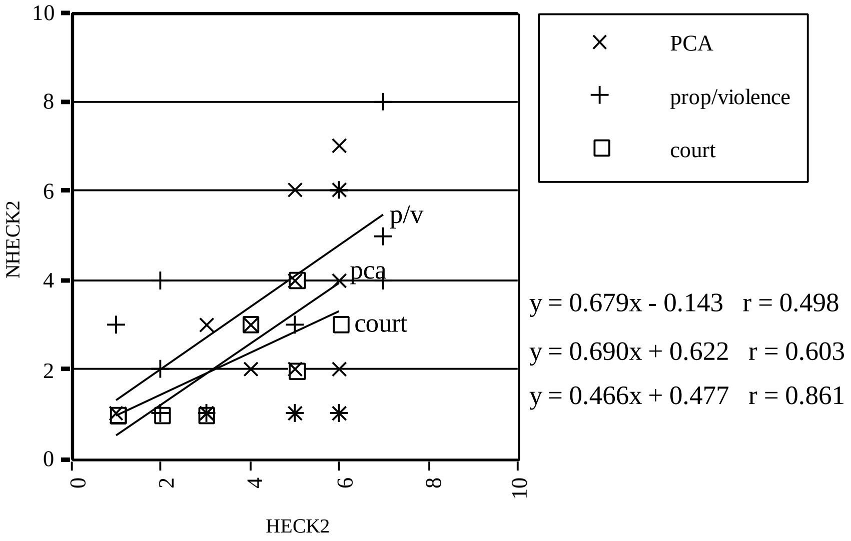


Heckling of Offender

40. How much moral lecturing was directed at the offender?

Overall correlation: .67
Agreement. overall: 58% PCA cases: 47% P/V cases 47% Court cases 80%.

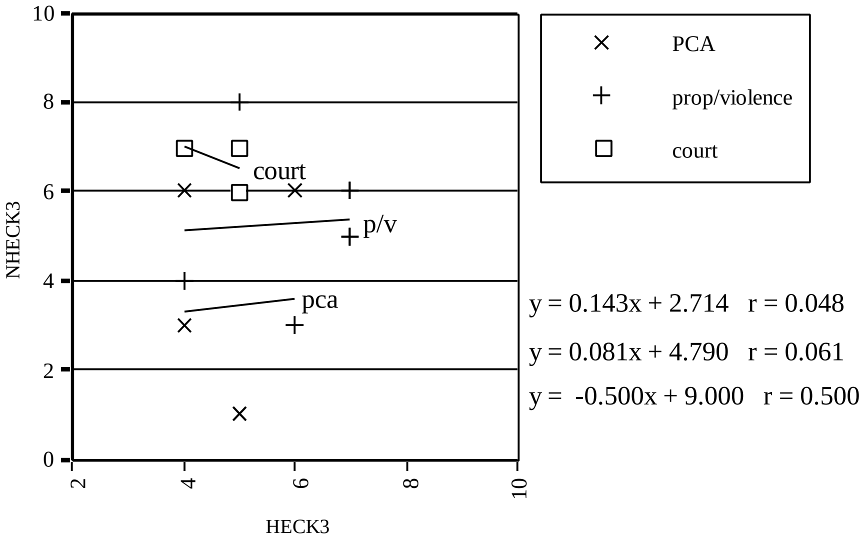
Comments: Agreement is low for the conference cases. Again there seem to be differences depending upon the raters. Observer 3 has higher agreement with PCA cases whereas observer 2 has higher agreement for P/V cases. There is fairly good agreement for court cases.



41. How clearly were the possible consequences of future offences communicated to the offender?

Overall correlation: .1
Agreement. overall: 40% PCA cases: - P/V cases -

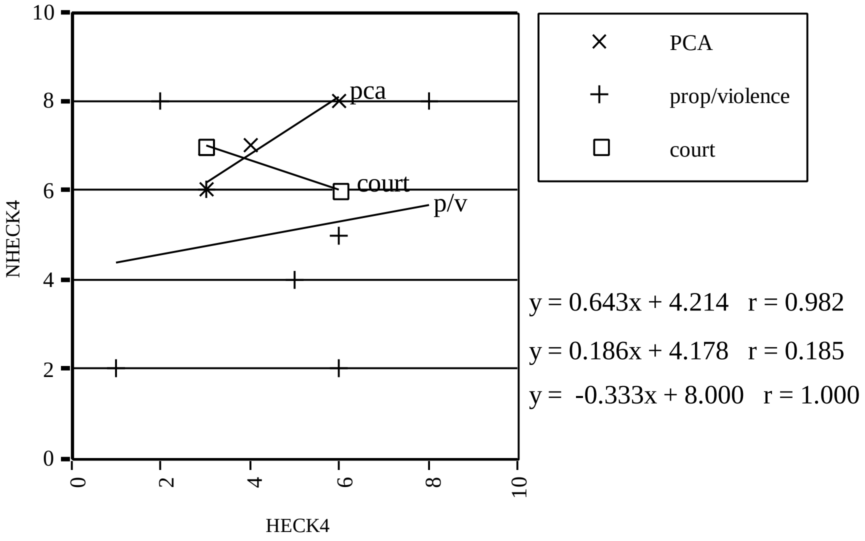
Comments: There are too few cases to test inter-rater agreement.



42. If the possible consequences of future offences were communicated to the offender, to what extent was this done in a non-threatening or matter-of-fact way?

Overall correlation: .13
Agreement. overall: 38% PCA cases: - P/V cases -

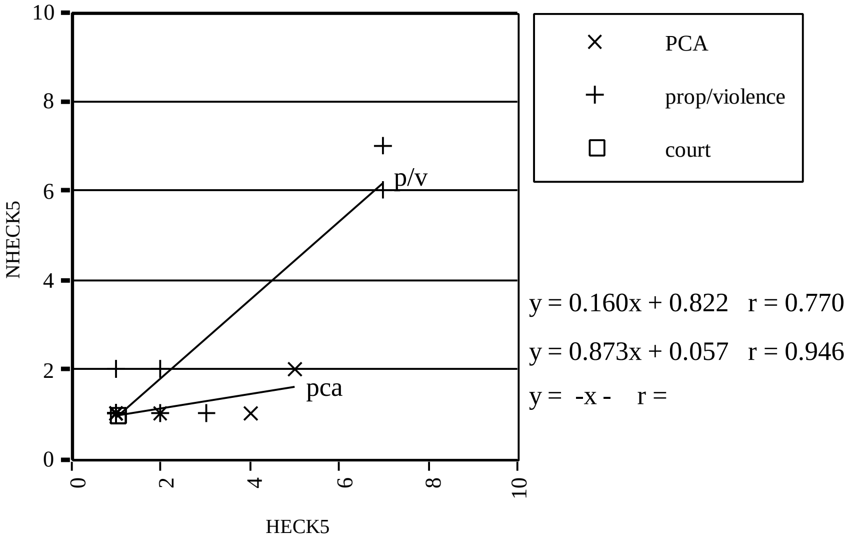
Comments: Again this category has not occurred enough to test inter-rater reliability. The total number of times the category occurred was 17.



43. How much was the offender harassed?

Overall correlation: .8
Agreement. overall: 93% PCA cases: 87% P/V cases 93% Court cases 100%.

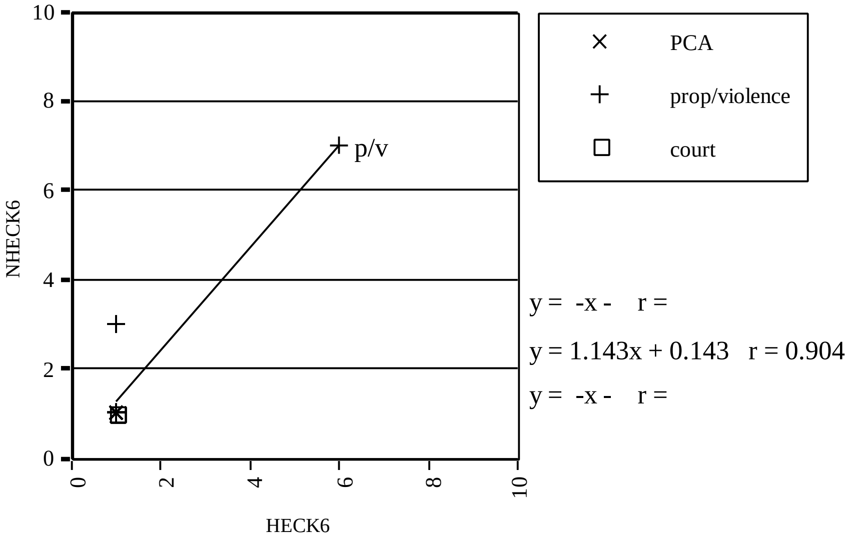
Comments: Again the variation for this question is limited, with none for court, which suggests that the results should be treated with caution. All contexts had high agreement scores.



44. How often was the offender shouted at?

Overall correlation: .9
Agreement. overall: 96% PCA cases: 100% P/V cases 87% Court cases 100%.

Comments: In all observation contexts the raters agreed that there is very little of this category but there is not enough variation to adequately test the inter-rater agreement.

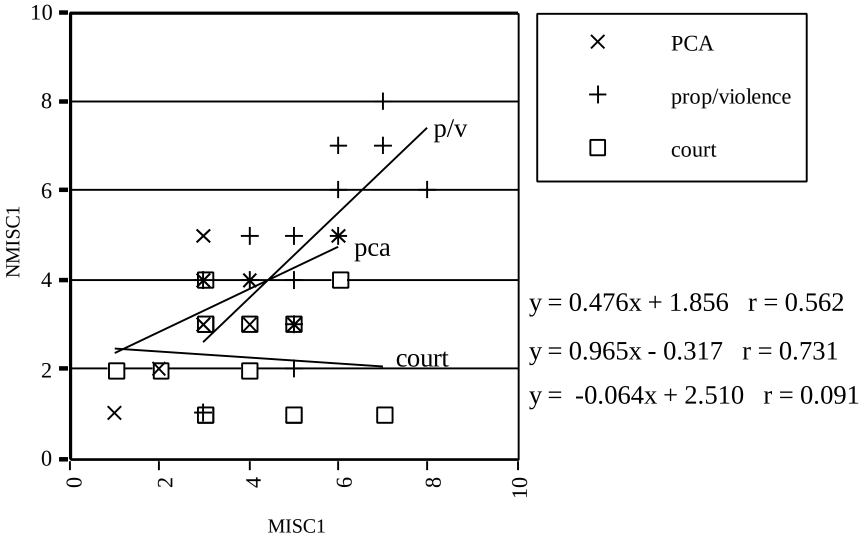


Miscellany

45. Overall how emotionally engaged was the offender?

Overall correlation: .53
Agreement. overall: 64% PCA cases: 73% P/V cases 71% Court cases 47%.

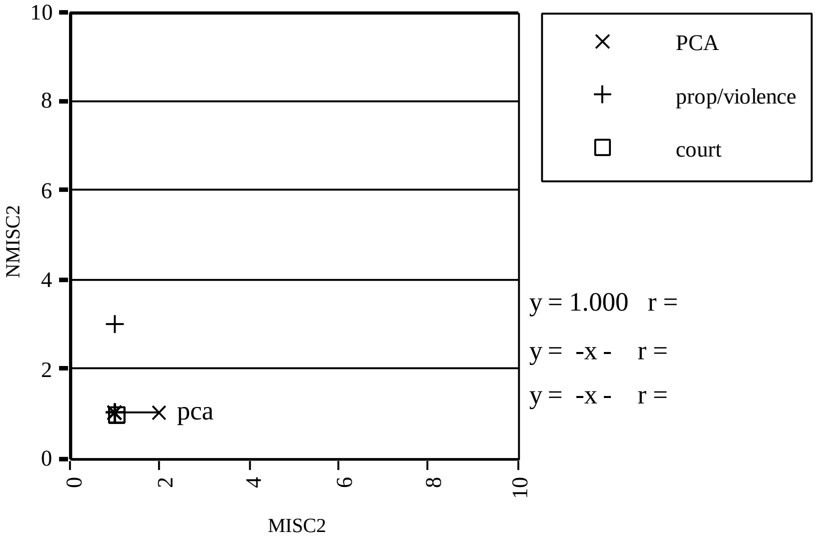
Comments: There is very good inter-rater agreement for the P/V cases, good agreement for the PCA cases and little for the court cases.



46. How much approval of the offender's criminal actions was expressed?

Overall correlation: -.02
Agreement. overall: 98% PCA cases: 100% P/V cases 93% Court cases 100%.

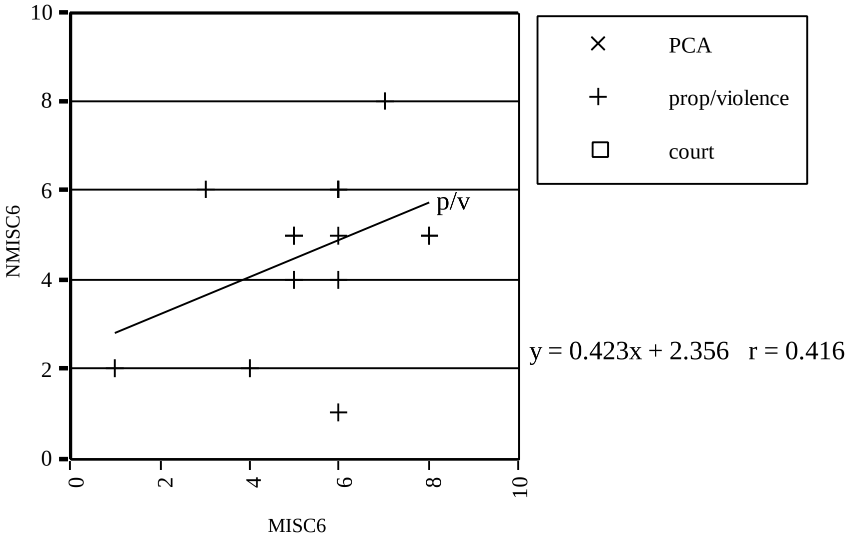
Comments: In all observation contexts the raters agreed that there was very little of this category but there is not enough variation to adequately test the inter-rater agreement.



47. How much discussion was there of the needs of the victim(s)?

Overall correlation: .42
Agreement. overall: 60% PCA cases: - P/V cases 60% Court cases -

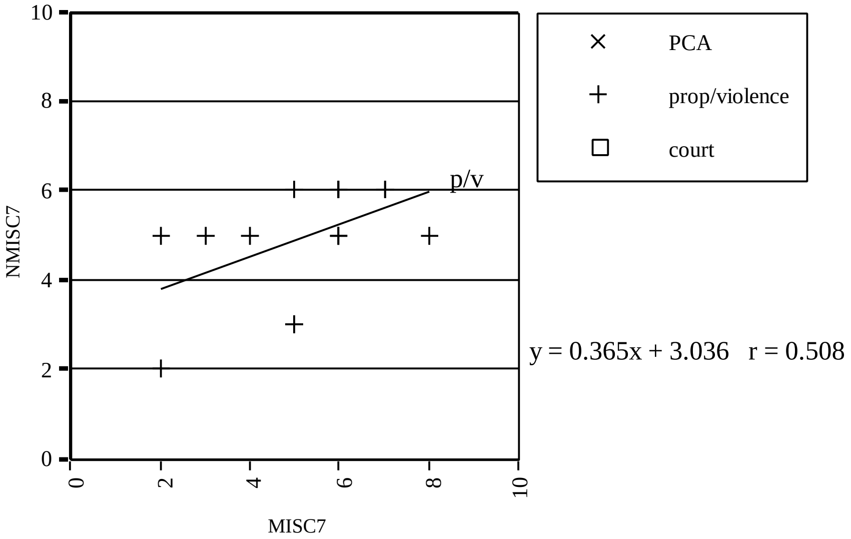
Comments: This question was only asked in P/V cases. There was much higher agreement for observer 2 (.71) than there was with observer 3 (.01).



48. How much support was given to the victim(s)?

Overall correlation: .51
Agreement. overall: 67% PCA cases: - P/V cases 67% Court cases -

Comments: Again asked only for P/V cases, this item has moderate inter-rater agreement.



One important pattern evident in the data is the degree of variation in agreement between observation contexts. Very few questions have high inter-rater agreement in all three contexts and most have poor agreement in at least one. This is quite important because it suggests that for some questions a more complete understanding of inter-rater agreement requires one to look at agreement for each context. For example, although inter-rater agreement across all cases is only 53% on question 4, it is 72% for court cases.

There are a number of possible reasons for this finding. One which seems quite likely is that some events are much easier to observe in some contexts than others. For example (question 28), the extent of discussion about the direct consequences of an offender's actions is much more easily observed in property or violence conferences (87% agreement) than in drink drive conferences (20% agreement). In most cases the difficulty seems to occur because a behaviour is much more obscure in one of the contexts; for instance, in the above example discussion of the specific consequences (ie those which actually resulted from the incident) of having been picked up at a random breath test are often obscured by the discussion of the general consequences of drink driving (eg road accidents).

A second possible reason for variability over contexts is differences between the observers' interpretation of questions. For example, in question 8, one observer has considered more general disapproval of the offence as an instance of disapproval of the offender's act whereas the other has not. Because the distinction between general and specific disapproval is much more distinct in property or violence cases there was quite high agreement (80%), however in drink drive conferences where disapproval is often more general, the observers agreed much less often (40%). This has important implications for training of interviewers.

Also important in interpretation of the data is the comparison between the two combination of observers. So as to determine how this affected the results, the agreement of each of these observers with rater 1 was assessed separately (see Appendix 3). What this analysis suggests is that for a number questions observer 2 and 3 varied in relation to rater 1 quite differently (eg questions 1, 2, 6, 25, 26, 34, 38, 41). However, this difference between observers 2 and 3 should be expected and is equivalent to the disagreement found between the two raters. While the difference between the combinations of observers increases the complexity of the data it is unlikely to alter the final conclusions. If anything the additional observer only serves to test inter-rater agreement more fully.

Despite the differences between observation contexts and the effect of different combinations of observers, the results suggest that there is high inter-rater agreement concerning a number of the concepts that the global observation instrument set out to measure.

The most central of these concepts is that of Reintegrative Shaming (Braithwaite 1989) which is measured most obviously in question 1 ("How much reintegrative shaming was expressed?"). The results show a high level of agreement between observers (see Appendix 2) which implies that it is possible for trained observers to reliably judge how much reintegrative shaming occurs in a case. While this is quite significant it will be of even more interest to see whether the observer's ratings are validated by the participant interviews.

Of further interest are the four facets which Makkai and Braithwaite (1994) identify as defining the differences between reintegrative and stigmatic shame. However, central to each of these facets is assumption that shaming occurred and so it was also necessary to measure whether there was 'disapproval of the act' by those present at the case. The scatterplots show that 'disapproval of the act' was measured with high agreement between raters in all conditions, except PCA conferences. Qualitative experience suggests, as discussed above, that in PCA conferences there were differences in observers' interpretation of question regarding disapproval of the type of offence and disapproval of the specific offence. This would seem a likely cause of lower agreement in this context and is an important issue for training observers.

The first facet outlined by Braithwaite involves disapproval while sustaining a relationship of respect which was successfully measured by a number of questions in the category 'respect for offender'. Although this facet was best measured by question 2 ('How much support was the offender given during the conference/court case?) all the questions have reasonable agreement in at least one of the observation conditions.

A second facet of reintegration is the termination of deviance which the instrument attempted to measure through a number of questions in the category ‘forgiveness of offender’. This facet was measured with moderate agreement between raters on only one of the three questions. While this is a poor result , it might have been expected as qualitative experience suggest that forgiveness is difficult to observe at the end of a conference because the process of forgiveness is a subtle one which often continues once the conferences is over. It is expected that forgiveness will be measured much better in the participant interviews.

The third (questions 9, 10, 13 and 14) and fourth facets (questions 11 and 12) were measured by the category ‘disapproval of offender’. In both cases raters had very high agreement that neither occurred very often in any of the observational contexts. However, apart from question 10, the range of scores in this category was so limited that this data provides an inadequate test of these questions. The nature of these facets is such that obvious examples of it occur rarely, suggesting that quite a large sample would be needed to assess inter-rater agreement properly.

A second concept included in the study was that of defiance by the offender (Sherman 1993). This concept encountered the same difficulties that ‘disapproval of the person’ and ‘heckling of the offender’ did. Variation in the sample was not great enough to properly assess inter-rater agreement. As with ‘disapproval of the offender’, these behaviours do not occur very often in conferences or court cases and thus with a relatively small sample of cases it is difficult to test them fully. However, it is encouraging that there was very high agreement that defiance rarely occurred.

A third concept that the instrument measured is the extent to which offenders displayed behaviour consistent with the emotion of shame. The results show quite good agreement between observers on a number of these questions across all three contexts, suggesting that behaviour associated with the emotion can be observed reliably.

Finally, the global observation instrument attempted to measure a number of elements central to the procedures involved in conferencing. Measured with reasonable inter-rater agreement were the categories ‘consequences of the offender’s act’ and ‘offender apologises’. However, measured less well were the categories ‘reparation to victim/community’ and ‘procedural justice’. Both of these areas had quite low agreement across contexts, but it is expected that these concepts can be better measured in the participant interviews.

These results suggest that a number of the concepts in this study can be measured reliably by the Global Ratings Questionnaire, despite differences between contexts. Such concepts include reintegrative shaming, shame, remorse by the offender and consequences of the act. A number of other concepts (defiance, disapproval of the offender, and heckling of the offender) could not be adequately tested because occurrences of the behaviour occurred so rarely in the sample. While greater variation in the sample would have been needed to properly assess these variables the results were encouraging in that there was very high inter-rater agreement that the behaviours rarely occurred. Finally a number of concepts (reparation to the community / victim, procedural justice and forgiveness) were measured with low inter-rater agreement.

It is evident from these results that the instrument can be improved in a number of ways. It would seem that a number of questions have lower agreement because of different interpretations by raters. So as to improve this a codebook which provides much better definition of questions has been written (see Appendix 7). In addition to interpretation problems it is also evident that a number of questions could be improved. The addition of new questions (see Appendix 5) in the areas of procedural justice and reparation will hopefully improve the measurement of these areas.

Systematic Observation Instrument

Although the primary purpose of these analyses was to investigate the extent of inter-rater agreement for the Systematic Observation Instrument, it was first necessary to conduct other more exploratory analyses to test assumptions about the experimental procedure and to explore the nature of the data collected.

An inspection of cell means indicated considerable variation between observation contexts (PCA conference, P/V conference, court case) and categories (respect, disapproval of act, disapproval of offender, apology, forgiveness, defiance, consequences, outcome) (see Table 2). In particular, Table 2 reveals that the frequency with which categories were used was very different, and also varied considerably across observation contexts. The following results are reported separately for observation contexts and categories in view of these sources of variation.

Table 2: Mean frequency of observations by observation context and category

	PCA (n=15)	P/V (n=15)	Court (n=15)
Respect for offender	3.07	3.07	0.87
Disapprove of act	2.8	4.2	0.53
Disapprove of offender	.73	1.6	0.07
Offender apologises	0	1.13	0.27
Offender is forgiven	0	0.20	0
Offender is defiant	0.40	0.26	0.07
Consequences	4.67	6.0	0.27
Outcome	1.67	2.87	0.80

It is clear from Table 2 that the frequency of observations among observational contexts is different, particularly between conference and court observations. This difference is undoubtedly due to the difference in treatment time; the five minute court case cannot possibly be as rich in relevant observational material as a 90 minute conference. However, it is also clear that a difference between PCA and P/V conferences exists. P/V conferences attracted more observations than PCA conferences in all categories except respect for the offender and defiance categories. It is likely that the presence of a direct victim at P/V conferences made these conferences more 'enriched'. Such victims may have expressed animosity toward the offender, being more disapproving and spending longer on discussion of consequences and potential outcomes. The presence of a direct victim would certainly prompt expressions of apology which were absent at PCA conferences.

Mean frequency of category use also showed that, overall, categories were differentially used. In particular, the categories of apology, forgiveness, and defiance were very much under-used. Of course, one does not expect multiple apologies and acts of forgiveness. However, given the instrument's design to measure events systematically, it is expected that these categories will be used with at least some frequency, and not never as was the case with some categories at PCA conferences and court. There are a number of possible reasons for the under-use of categories like apology, forgiveness, and defiance. Such categories are much less explicit than others which focus on discussion of a particular topic, discussion of the consequences or of the outcome, for example. The former categories have a behavioural rather than a verbal expression. Behavioural expressions are more likely to be tainted by the observer's perspective and expectation of events. These categories also have a implicit character in that their expression is part of the overall ambience of interactions and relative to previous interactions. This means there is often no observable and codeable chunk of behaviour representing expression of forgiveness or defiance that may be measured systematically. It is also true that apology and forgiveness may occur after completion of the case when observations have ceased. For these reasons it is suggested that these concepts might be better measured via interviews with participants.

It will be recalled that there was a change in the observer acting as the second rater two thirds the way through the data collection phase. It was therefore necessary to demonstrate that the two observers serving as rater two performed in an acceptably similar manner, before proceeding to establish whether rater one and two were in agreement. This was done by using rater one's observations as a constant and correlating them with the observations of the two respective observers acting as rater two. Correlations between observer 1 and 2 and between 1 and 3 were then calculated for each category to determine comparative inter-observer agreement (see Table 3).

Observations in which both observers made no observation for the category in question were removed from the analysis. Accordingly, some of the correlations should be treated with caution because of the small sample sizes.

The correlations between observer 1 and 2 and 1 and 3 on the respect for offender, disapprove of act, consequences and outcome categories are remarkably similar. Correlations for the disapproval of the offender category are rather more discrepant, but are based on an inadequate sample size. The remaining categories occurred so infrequently in the sample that it is difficult to properly assess the extent of inter-rater agreement, however it is evident that raters agree concerning the infrequency of the categories occurrence. These congruent correlations between sets of observers provide support that certain categories on the systematic observation instrument are reliable.

Table 3: Inter-observer agreement regarding the total number of units observed for observer 1 with observers 2 and 3.

	Obs 1 with Obs 2	Obs 1 with Obs 3
Respect for offender	0.72** (n=13)	0.72** (n=13)
Disapprove of act	0.70** (n=14)	0.62* (n=14)
Disapprove of offender	0.63 (n=7)	-0.58 (n=6)
Offender apologises	- (n=0)	0.54 (n=5)
Offender is forgiven	- (n=0)	- (n=3)
Offender is defiant	- (n=3)	- (n=2)
Consequences	0.79** (n=14)	0.78** (n=14)
Outcome	0.73** (n=14)	0.81** (n=14)
Total	0.84** (n=14)	0.86** (n=14)

**p<.01
*p<.05

Although there are clearly problems with determining agreement for those categories with a small sample size, the degree of corresponding agreement on less problematic categories was considered more than sufficient to support the assumption that observers 2 and 3 could be treated as one. Having demonstrated this, inter-rater agreements could be investigated.

The extent of inter-rater agreement with respect to the total number of units coded across observations was measured using Pearson correlation coefficients. Inter-rater agreement coefficients were calculated for each systematic observation category and separately for each observation context (see Table 4). The high frequency of observations with zero codes for both observers had a tendency to positively skew category distributions, thus causing correlations to be spuriously high. This problem was addressed by removing these cases before calculating coefficients. However, it should be noted that the analysis is conservative because it ignores a legitimate form of agreement; agreement that a category did not occur. Coefficient sample sizes are also reported because some of the sample sizes are inadequate.

It is important to note that the correlations reported in Table 4 were based on proportions rather than frequencies. Proportions were preferred because they are an internally-derived measure (being based on the observer's own total number of observations), making individual observations comparable among each other and across observation contexts. In this sense, proportions serve to standardise units, which may well be important for future analyses. For those interested the correlations based upon frequency are included in appendix 4.

Table 4: Inter-rater agreement (Pearson correlations) for categories across observation contexts.

	Total (N=45)	PCA (n=15)	P/V (n=15)	COURT (15)
Respect for offender	0.86** (n=37)	0.74** (n=14)	0.74** (n=14)	0.79** (n=9)
Disapprove of act	0.68** (n=37)	0.67** (n=15)	0.48 (n=15)	0.67 (n=7)
Disapprove of offender	0.11 (n=22)	0.63 (n=7)	0.30 (n=12)	-0.50 (n=3)
Offender apologises	0.82** (n=15)	-- (n=0)	0.50 (n=12)	0.99 (n=3)
Offender is forgiven	-0.22 (n=4)	-- (n=0)	-0.22 (n=4)	-- (n=0)
Offender is defiant	0.17 (n=9)	0.47 (n=4)	0.02 (n=4)	-- (n=1)
Consequences	0.62** (n=35)	0.80** (n=15	0.49 (n=15)	0.42 (n=5)
Outcome	0.73** (n=43)	0.75** (n=15)	0.52* (n=15)	0.63* (n=13

**p<.01
*p<.05

Overall, inter-rater agreement was high for the respect for the offender, disapproval of the act, consequences, offender apologises, and outcome categories. Inter-rater agreements for the other categories were again problematic because they occurred so rarely. As with the Global Ratings Questionnaire, these concepts were reliably recorded by observers as having not occurred. This however does not provide an adequate test of these categories which would require a larger sample size. The results for the disapproval of the offender category are disappointing but not surprising given the discrepancy in the way observers 1 and 2 and observers 1 and 3 made observations relevant to this category (see Table 3).

The Systematic Observation Instrument was designed to provide information about the sequence of events at court and conferences. Does 'respect for the offender' usually follow instances of 'disapproving of the act', for example? Thus, it was important to demonstrate that the frequency of such sequences were reliably observed. The reliability of several sequences, those with sufficient data, were investigated across all observations and for each observation context separately (see Table 5). This was done by calculating the frequency with which consecutive codes matched the sequence under investigation and correlating sequence frequencies between observers across observations. Again, to avoid spuriously high correlations due to neither observer recorded the sequence investigated such observations were removed.

Table 5: Inter-rater agreement (Pearson's correlations) regarding the frequency with which various category sequences occur.

	Total (n=45)	PCA (n=15)	P/V (n=15)	Court (n=15)
Disapproval of act followed by Consequences	0.70** (n=31)	0.82** (n=14)	0.58* (n=15)	- (n=2)
Disapproval of act followed by Respect	-0.09 (n=17)	0 (n=9)	-.26 (n=7)	- (n=1)
Consequences followed by Outcome	0.46* (n=27)	0.50 (n=14)	0.41 (n=12)	- (n=1)
Consequences followed by Respect	0.71** (n=31)	0.61* (n=14)	0.87** (n=13)	-0.58 (n=4)
Respect followed by Disapproval of act followed by Consequences	0.49 (n=14)	0.68 (n=6)	-0.65 (n=5)	- (n=3)

**p<.01
*p<.05

Inter-rater agreement regarding the frequency with which various category sequences occur further support the reliability of the systematic observation instrument. Although the two raters' observations were not matched, observation for observation, it was still possible to demonstrate that observers can agree about the order of certain observations.

The results from the Systematic Observation Instrument demonstrate that observations are made differently depending on the observation context. Moreover, the frequency with which categories are used varies markedly. However, while noting these features of the Systematic Observation Instrument, it was also evident that the categories ‘respect for the offender’, ‘disapproval of the act’, ‘consequences’ and ‘outcome’ categories were reliable. The consistency of inter-rater agreements between two sets of observers also suggests that the present training regime is capable of producing observers with a standardised frame of reference. The fair inter-rater agreement for observation sequences demonstrates that the order of events occurring at conferences and court can also be recorded in a like way by different observers. Similarly, while it is true that reliability has not been demonstrated for apology, forgiveness, and defiance categories, it is also true that these categories have not as yet been given a fair test due to insufficient samples sizes.

The results for the category of ‘disapprove of the offender’ suggests that there were differences in interpretation between observers (see tables 3 and 4). While one combination of observers (1 and 2) appeared to have reasonable agreement the second combination (1 and 3) demonstrated no agreement at all. This would suggest that the poor results are probably due to differences in interpretation of categories between observers. Qualitative experience suggests that there was some confusion about the categories of disapproval of the offender and consequences of the act which most likely accounts for the poor results. It is important that in future training the definition of these categories is emphasised. Equally, so as to clarify these categories, changes to the Systematic Observation Codebook have been made (see Appendix 6).

Conclusions

The results of this study suggest that a number of the concepts central to Diversionary Conferences can be measured reliably on both the Global Ratings Questionnaire and the Systematic Observation Instrument. Particularly encouraging was the reliable measurement of reintegrative shaming, shame, consequences of the act, and the ‘outcome’ category. The results also highlighted the need for a number of improvements in the instruments. Disagreement between observers due of differences in the interpretation of questions highlights the need for better definitions of items and a greater emphasis on training. Equally, because a number of areas in the Global Rating Questionnaire were weak a number of new questions were drafted (see Appendix 5).

References

1. Braithwaite, J. (1989), Crime, Shame and Reintegration. Cambridge: Cambridge University Press.
2. Makkai, T. and Braithwaite, J. (1994), 'Reintegrative Shaming and Compliance with Regulatory Standards', Criminology, 32, 3, 361-385.
3. Scheff, T. J., and Retzinger, S. M. (1991), Emotions and Violence: Shame and Rage in Destructive Conflicts. Lexington, MA: Lexington Books.
4. Sherman, L. W. (1993), Defiance, Deterrence and Irrelevance: A Theory of the Criminal Sanction. Journal of Research in Crime and Delinquency, 30, 4, 445-473.
5. Sherman, L. W., Braithwaite, J. and Strang, H. (1994), Reintegrative Shaming of Violence, Drink Driving and Property Crime: A Randomised Controlled Trial. Unpublished Grant Proposal.