Using Cohen's Kappa to Calculate Inter-Rater Reliability

There is a problem with using simple percentages to assess inter-rater reliability. Suppose two raters watched ten segments of TV, looking for examples of aggressive body language (for example) and came up with the following (Y=Yes, N=No):

Segment	1	2	3	4	5	6	7	8	9	10
Rater A	Y	Y	N	Y	Y	Y	N	Y	Y	Y
Rater B	Y	Y	N	N	Y	N	Y	Y	Y	Y

In other words, the two raters agree in seven of the ten segments. 70% sounds like reasonable agreement, until you see the problem: no account is taken of the amount of agreement predicted by chance.

Note that Rater A has said Yes in 80% of the segments, and Rater B has said Yes in 70% of the segments. This means that, on average, they will agree with Yes in 56% of segments (that's 0.8 x 0.7), and agree with No in 6% of segments (that's 0.2 x 0.3), giving a total agreement of 62% regardless of what they are actually measuring. Suddenly, 70% does not sound like such a good score compared to 62% predicted by chance!

A better measure of reliability takes this into account. Cohen's Kappa (k) is calculated as follows:

$$k = (Po - Pc) / (1 - Pc)$$

where Po is the observed proportion of agreement and Pc is the proportion predicted by chance. In this example:

$$k = (0.70 - 0.62) / (1- 0.62) = 0.21$$

A good rule of thumb is that Kappa should be 0.7 or more, so in this example there is clearly a problem with the category. Our researchers must go back to the drawing board!

Mike Eslea, 7/10/96