5.6 Agreement on Culposcopic Photographs for Child Sexual Abuse

4. A brave group of investigators{Sinal, 1997 #204} examined inter-rater reliability of clinicians interpreting culposcopic photographs for the diagnosis of sexual abuse in prepubertal girls. Experienced clinicians (N = 7) rated sets of photographs on the following 5-point scale: 1, normal; 2, nonspecific findings; 3, suspicious for abuse; 4, suggestive of penetration; 5, clear evidence of penetration.

   a) The published unweighted kappa in this study was 0.20; the published weighted kappa (using quadratic weights) was 0.62. Why do you think there is a big difference between them?

b.) The authors used quadratic weights. As shown in Table 5.5, these weights give 43.75% credit for answers that are 3 categories apart (e.g., "normal" and "suggestive of penetration." This might seem excessively generous. Propose an alternative weighting scheme, by creating a 5 x 5 table with weights (you only need to include the numbers above the diagonal) and justify it. (Hint: Don't just use linear-weighted Kappa. Ask yourself: are some 1-level disagreements more clinically significant than others? Should there be any credit at all for 3-level disagreements?)

   c) The data collection form for the study included a sixth category: "unable to interpret." Most of the kappa values published for the study were based on the subset of 77 (55%) of 139 sets of photographs that were "interpretable" by all 7 clinicians.

      i. Did including an "unable to interpret" category and then excluding photographs for which anyone selected that category probably increase or decrease kappa (compared with not including that category)?

      ii. How else could they have handled that sixth "unable to interpret" category?

   d) The practitioners who participated in this study were all trained in evaluating suspected sexual abuse, with a minimum experience of 50 previous cases (6 of 7 had seen more than 100 previous cases). How does this affect the generalizability of the results and your conclusions?

e) The authors actually assessed inter-observer agreement in two groups of clinicians, both with and without blinding them to the patients' histories. Results are shown below:

(Unweighted) Kappa Values for Interpretation of Culposcopic Photos on a 5-Point Scale

	Blinded (N = 456)[a]	Provided History (N = 510)[a]
Group 1	0.22	0.11
Group 2	0.31	0.15

[a] These N values indicate the number of pairwise comparisons in which both clinicians considered the photograph to be interpretable.

What are some possible explanations for the higher kappa values when observers were blinded to the history?