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?

Weighted Kappa gives partial credit for being close, whereas unweighted Kappa counts only perfect agreement along the diagonal. If observers almost never completely disagree (in this case one observer saying "normal" and the other saying "clear evidence of penetration") weighted Kappa will generally be higher than unweighted Kappa, and if most disagreements are only a category or two, weighted Kappa will be much higher, especially using quadratic weights (see part b).

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?)

Here is one set of custom weights.

(1, normal; 2, nonspecific findings; 3, suspicious for abuse; 4, suggestive of penetration; 5, clear evidence of penetration).

	1	2	3	4	5
1	**1**	**.75**	**0**	**0**	**0**
2		**1**	**.1**	**0**	**0**
3			**1**	**.5**	**0**
4				**1**	**.5**
5					**1**

This weighting scheme treats "normal" and "non-specific" as near agreement. It gives half credit if one observer says "suspicious for abuse" and another says "suggestive of penetration," because those seem similar to us. It also gives half credit for "suggestive of penetration" and "clear evidence of penetration." But since

the clinical implications of "nonspecific" and "suggestive of abuse" seem very different, it does not provide much credit for that disagreement, and there's no credit at all for any answers that are 2 or more categories apart.

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)?

The exclusion would probably increase kappa by limiting the comparison only to photos that all raters agreed were "interpretable." If forced to interpret photos they believe to be uninterpretable, the clinicians looking at the photos would need to guess.

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

They could have included a 6th category in the grid, for "unable to interpret," to see if the raters agreed on that rating. This would have precluded use of weighted kappa, however, unless they could place "unable to interpret" on the ordinal scale of the findings. Alternatively, they could have combined "unable to interpret" with "nonspecific findings" – in both cases the rater is making no judgment about sexual abuse − which would preserve the ability to calculate weighted kappa.

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?

The estimates of kappa from this study are probably higher than would be obtained with less experienced examiners. If the conclusion of the study is that inter-rater reliability is not very good, this would only be strengthened by the high level of experience of the examiners. On the other hand, although it seems unlikely in this setting, it is worth at least considering the possibility that they see a referral population in whom findings are especially difficult to interpret, in which case Kappa could be falsely low.

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?

This is a fascinating and counter-intuitive finding. One would expect kappa to increase with provision of more information. The drop in kappa is probably due to some combination of:

1. **Interobserver agreement on interpretation of the history is worse than agreement on physical findings. The lower kappa when history is provided suggests**

   a) that they are using the history to interpret the physical examination, and

   b) they disagree about how to do this.

2. **The [a] footnotes indicate that the sample size was higher when the history was provided, presumably because fewer photographs were regarded as uninterpretable. Perhaps the agreement on these photos was very poor.**

3. **The difference could be due to chance. Confidence intervals are not provided, but given the sample size and the consistency and magnitude of the difference, it seems chance is probably not the whole explanation.**

4. **The authors made a mistake in analyzing or publishing their results.**

Note: Some of our students have suggested that if the history increased the agreement on the marginals, this would increase the expected agreement, and could therefore lead to a decrease in kappa. However, we can't think of any mechanism by which telling clinicians the history associated with each photo would lead to greater agreement on the marginals without correspondingly greater agreement within the table, which would tend to increase rather than decrease Kappa.