INTEROBSERVER VARIABILITY IN GRADING
PERTHES' DISEASE

E. D. SIMMONS, H. K. GRAHAM, J. P. SZALAI

From the Hospital for Sick Children, Toronto

Fifteen independent observers of three levels of experience (consultant staff, fellows, residents) assessed 40 radiographs of children presenting with Perthes' disease using the Catterall and the Salter-Thompson grading systems. Each observer was supplied with descriptions and illustrations of the classifications and each hip was grouped by both systems by each observer. The results were statistically analysed using 'kappa' statistics.

The level of interobserver agreement was higher for the Salter-Thompson system and correlated with the level of experience of the observer. Both systems can give acceptable levels of interobserver agreement, but the Salter-Thompson grouping is simpler and easier to apply in the earlier stages of the disease when treatment must be decided, and has a higher degree of reproducibility amongst more experienced observers.

Current understanding of the natural history of Perthes' disease is largely based on radiographic studies. Various classifications of radiographic appearances have been developed, and are key factors in selecting treatment (Figs 1 to 4). Waldenström (1922) described four sequential radiographic phases of the disease and was probably the first to note partial involvement of the femoral capital epiphysis (1938).

O'Garra (1959) developed the concept of partial and complete involvement of the epiphysis, describing two distinct groups of patients with Perthes' disease. His group 1 patients had mainly anterior involvement of the epiphysis, seen on the lateral radiograph, and a good prognosis. His group 2 patients had involvement of the whole epiphysis, a more protracted course, and a worse prognosis. Catterall (1971) divided these two groups into a four-group classification which has since been used by many authors (Kamhi and MacEwen 1975; Lloyd-Roberts, Catterall and Salamon 1976; Brotherton and McKibbin 1977; Dickens and Menelaus 1978; Blake-more and Harrison 1979; Kelly, Canale and Jones 1980; Stevens, Williams and Menelaus 1981; Van Dam et al 1981; Wenger 1981).

With experience, a number of criticisms of the Catterall grouping have been made. First, since the classification is based on the epiphyseal involvement at the time of maximum resorption and this occurs relatively late, the group will appear to change if the classification is made too early (Kelly et al 1980; Van Dam et al 1981; Wenber 1981). Secondly, it can be difficult to apply the classification. The only two studies which have addressed this question both report an unacceptably poor level of interobserver agreement (Hardcastle et al 1980; Christensen et al 1986).

In 1984, Salter and Thompson proposed a return to a simpler two-group classification based on the extent of the subchondral fracture. This classification may be applied earlier in the course of disease and is being used to select patients for various management options in some prospective studies of Perthes' disease. We have now assessed interobserver variability in the use of the Salter-Thompson system, and re-assessed the Catterall system.

MATERIALS AND METHODS

Good quality anteroposterior and frog lateral radiographs of 40 children who presented at the Hospital for Sick Children, Toronto, between 1982 and 1984 were chosen. They were selected for good visual quality from those taken at the time of first presentation for treatment. The chosen radiographs were photographed under standard conditions and made into slides. The slides were then
shown to 15 observers of three levels of experience: five attending staff surgeons (consultants), five postgraduate clinical fellows, and five senior residents. All observers were familiar with both Catterall and Salter–Thompson grading systems, and were, in addition, provided with descriptions and diagrams of both systems.

Each of the 40 hips was graded with both systems by each observer independently, choices being marked on special sheets. Ample time was allowed for each observer to make his decision on each hip, there was no communication between observers. The results were analysed using ‘kappa’ statistics to calculate weighted coefficients for the sets of results.

Weighted ‘kappa’ coefficients are a measure of interobserver agreement beyond that which would be expected by chance alone. The level of agreement and the ‘seriousness’ of the disagreement are assessed by a weighted formula, which provides more accuracy than percentage agreement alone (Cohen 1968; Light 1971; Fleiss 1981). As an example, kappa statistics can distinguish between a pair of observers, one of whom chooses Catterall group 1 and group 2 for one hip, and another pair of observers choosing Catterall group 1 and group 4 for the same hip. With simple percentage agreement statistics both differences are scored equivalently but kappa statistics will penalise the minor discrepancy less than the larger one.

A ‘kappa’ coefficient is expressed on a scale from zero to one: zero indicates agreement only as would occur by chance alone, one indicates perfect agreement. Kappa values from 0 to 0.40 represent poor agreement, 0.4 to 0.75 fair to good agreement, and values of 0.75 to 1.0 excellent agreement (Fleiss 1981). The overall kappa values were calculated for each group of observers and for each classification.

RESULTS
The results of the study are shown in Tables I and II, which give overall weighted kappa values with standard errors and confidence limits. For comparison, percentage agreement figures are also given.

The kappa values and their confidence limits suggest that agreement was clearly above that which would have been expected to occur by chance. It also seems that the level of interobserver agreement may be dependent on the level of experience of the observing group, though all participants were provided with reference material. This trend is more noticeable with the Salter–Thompson classification than with the Catterall classification.

In comparing the Catterall and Salter–Thompson systems, there is a higher level of interobserver agreement for the most experienced group (consultant staff) when using the Salter–Thompson system, but no significant
difference between the two classification systems for the two less experienced groups (fellows, residents). The level of interobserver agreement is acceptable for both classification systems, but is highest for more experienced observers using the Salter–Thompson groups.

**DISCUSSION**

A classification system should be relatively easy to apply, should give information relevant to prognosis and/or treatment, and should have an acceptable level of reproducibility and reliability. Few classification systems have actually been tested for reproducibility between different observers. Two previous studies of the Catterall grouping have found that it failed to provide an acceptable level of interobserver agreement (Hardcastle et al 1980; Christensen et al 1986), but both studies had some methodological weaknesses. One used percentage agreement statistics only (Hardcastle et al 1980), and the other used only four observers and failed to report probability values for chance agreements (Christensen et al 1986). In neither study were the observers supplied with diagrams or descriptions of the classification.

Our study of the two systems, (to our knowledge ours is the first evaluation of the Salter–Thompson system) showed that both gave acceptable levels of interobserver agreement, though the Salter–Thompson yielded a higher degree of interobserver agreement when used by experienced observers. The Salter–Thompson system is simpler to use and can be applied early in the disease process – two factors in its favour. It has also been demonstrated that a two-group classification is all that is required for treatment and to gauge prognosis, and that there is no particular advantage in using a more complex four-group classification (Salter and Thompson 1984).

The authors wish to thank the orthopaedic staff of the Hospital for Sick Children, Toronto, for participating in this study.

No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

**REFERENCES**


Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968; 70: 213-20.


