Inter-observer reliability of radiographic classifications and measurements in the assessment of Perthes' disease

Ola Wiig[1], Terje Terjesen[2] and Svein Svenningsen[1]

Departments of [1]Orthopaedics, Aust–Agder Hospital, NO–4809 Arendal, Norway, [2]Orthopaedics, The National Hospital, NO–0570 Oslo, Norway. ola.wiig@aassh.no
Submitted 01–06–11. Accepted 02–04–02

ABSTRACT – We evaluated the inter-observer agreement of radiographic methods when evaluating patients with Perthes' disease. The radiographs were assessed at the time of diagnosis and at the 1-year follow-up by local orthopaedic surgeons (O) and 2 experienced pediatric orthopedic surgeons (TT and SS). The Catterall, Salter-Thompson, and Herring lateral pillar classifications were compared, and the femoral head coverage (FHC), center-edge angle (CE-angle), and articulo–trochanteric distance (ATD) were measured in the affected and normal hips.

On the primary evaluation, the lateral pillar and Salter-Thompson classifications had a higher level of agreement among the observers than the Catterall classification, but none of the classifications showed good agreement (weighted kappa values between O and SS 0.56, 0.54, 0.49, respectively). Combining Catterall groups 1 and 2 into one group, and groups 3 and 4 into another resulted in better agreement (kappa 0.55) than with the original 4-group system. The agreement was also better (kappa 0.62–0.70) between experienced than between less experienced examiners for all classifications.

The femoral head coverage was a more reliable and accurate measure than the CE-angle for quantifying the acetabular covering of the femoral head, as indicated by higher intraclass correlation coefficients (ICC) and smaller inter-observer differences. The ATD showed good agreement in all comparisons and had low inter-observer differences.

We conclude that all classifications of femoral head involvement are adequate in clinical work if the radiographic assessment is done by experienced examiners. When they are less experienced examiners, a 2-group classification or the lateral pillar classification is more reliable. For evaluation of containment of the femoral head, FHC is more appropriate than the CE-angle.

■

Several radiographic classifications have been used to establish an accurate diagnosis and prognosis in Perthes' disease. Catterall (1971) described a 4-group classification, simplified by Salter and Thompson (1984), which divides the hips into 2 groups. In 1992, Herring et al. presented a classification based on the degree of resorption of the lateral femoral head pillar during the fragmentation phase.

Several authors have reported poor inter-observer reliability with the Catterall classification (Christensen et al. 1978, Hardcastle et al. 1980, Ritterbush et al. 1993). Better inter-observer agreement has been found with the Salter-Thompson or the lateral pillar classification (Simmons et al. 1990, Herring et al. 1992, Ritterbush et al. 1993, Podeszwa et al. 2000). However, the results of these reports have differed considerably and we have found no inter-observer study which has compared the 3 classifications.

In the radiographic evaluation of Perthes' disease, the degree of containment is important. This can be measured by the center-edge (CE) angle of Wiberg (1939) or by femoral head coverage (Heyman and Herndon 1950). The articulo-trochanteric distance (ATD) is another relevant parameter, since it describes the degree of trochanteric overgrowth and length of the femoral neck

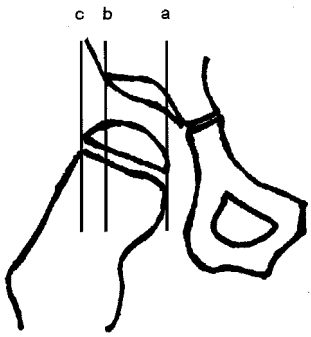

Figure 1. Femoral head coverage (FHC) is the percentage of the femoral head medial to Perkins' line (a–b) in relation to the width of the femoral head (a–c) (FHC = ab/ac × 100).

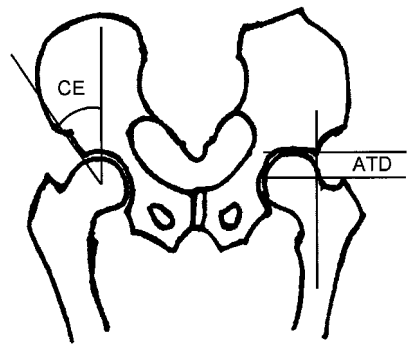

Figure 2. Schematic drawing showing the CE-angle (right hip) and the ATD (left hip).

(Edgren 1965). We assessed the inter-observer reliability of the radiographic classifications and measurements to determine which radiographic methods are most appropriate and reliable in the Perthes' hip.

Patients and methods

As part of a nationwide study on Perthes' disease, the primary radiographs and the radiographs taken 1 year after diagnosis were examined by orthopedic surgeons (O) with a special interest in pediatric orthopedics at the local hospitals where the patients were referred. These orthopedic surgeons were briefed on the theoretical basis and practical use of the classifications and measurements. The radiographs were then sent to a pediatric orthopedic surgeon with great experience in examining radiographs of hips in children (SS). He assessed all of them in order to make the evaluation as reliable and consistent as possible. The radiographs of every other patient (alphabetically) from the first 2 years of the study were also seen by an experienced pediatric orthopedic surgeon (TT).

The radiographic phases were determined at the time of diagnosis, when the initial phase was characterized by differences in epiphyseal height, width and bone structure from the normal hip. In the fragmentation phase, the necrotic bone in the hips was partly or totally reabsorbed. In the reossification phase, the hips showed obvious signs of reossification.

On the basis of anteroposterior and Lauenstein projections, the diseased hips were divided into 4 groups, as originally proposed by Catterall (1971), and into 2 groups as recomended by Salter and Thompson (1984), in which group A included hips with less than 50% involvement of the femoral head and group B those with more than 50%.

We also used the Herring classification (1992), in which the affected hips were divided into 3 groups: group A with no height reduction in the lateral pillar of the femoral head, group B, with more than 50% height of the lateral pillar maintained, and group C, with less than 50% height maintained.

The femoral head coverage was determined by calculating the percentage of the femoral head medial to Perkins' line in relation to the width of the femoral head parallel to Hilgenreiner's line (Figure 1) (described by Heyman and Herndon in 1950 as the acetabulum-head quotient).

The CE-angle (Wiberg 1939) is the angle between the perpendicular through the center of the femoral head and the line connecting the center of the femoral head and the lateral edge of the acetabulum (Figure 2). We measured this angle with Müller's method (1971).

The articulo-trochanteric distance (ATD) was measured as the distance between 2 lines perpendicular to Perkins' line: the line through the proximal tip of the greater trochanter and that through the most proximal point of the femoral head (Figure 2).

The categorical data were analyzed by weighted kappa statistics (Altmann 1999). We also determined the percentage agreement between the observers. The kappa has a maximum of 1 when

Table 1. Inter-observer agreement, Catterall classification

		Agreement						Kappa$_{(W)}$
		Catterall grouping						
	N_1	1	2	3	4	N_2	%	
Primary								
O/SS	158	8	11	38	43	100	63	0.49
TT/SS	76	2	5	35	15	57	75	0.62
Follow–up								
O/SS	115	2	1	12	51	66	57	0.28
TT/SS	63	–	–	20	26	46	73	–

O local orthopedic surgeons, SS and TT pediatric orthopedic surgeons, Primary radiographs at time of diagnosis, Follow-up radiographs at 1-year follow-up after diagnosis, N_1 number of patients examined, N_2 number of patients agreed upon, % percentage agreement, Kappa$_{(W)}$ weighted kappa.

Table 2. Inter-observer agreement, Salter-Thompson classification

		Agreement				Kappa$_{(W)}$
		Salter-Thompson groups				
	N_1	A	B	N_2	%	
Primary						
O/SS	149	18	110	128	86	0.54
TT/SS	73	9	56	65	89	0.63
Follow-up						
O/SS	91	2	81	83	91	0.29
TT/SS	63	1	55	56	89	0.18

For symbols, see Table 1.

agreement is perfect, but a value of 0 indicates no agreement better than chance, and negative values show worse than chance agreement. As suggested by Altmann, we interpreted the kappa values as follows: below 0.20 as poor agreement, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as good, and over 0.80 as very good agreement.

The numerical data (femoral head coverage, CE-angle, and ATD) were analyzed by calculating intraclass correlation coefficients (ICC), a measure of the proportion of variance that is attributable to individuals (McGraw and Wong 1996). An ICC of 1 indicates perfect agreement. Mean and standard deviation (SD) of the differences between observers and the range, which includes 95% of the inter-observer differences (mean ± 2SD) (Altmann 1999), were also calculated.

Results

At the time of diagnosis, 45% of the hips were classified by SS as being in the initial phase, 51% in the fragmentation phase and 4% in the reossification phase. The inter-observer agreements between O and SS and TT and SS were moderate (kappa 0.59 and 0.52).

Catterall classification (Table 1)

Using the Catterall classification at the time of diagnosis, we obtained a weighted kappa value of 0.49 between the local orthopedic surgeons (O) and SS (moderate agreement), and 0.62 between the more experienced observers (TT and SS) (good agreement). The agreement was slightly better between O and SS in the fragmentation phase (kappa 0.44, agreement 67%) than in the initial phase (kappa 0.40, agreement 48%).

At the 1-year follow-up, the kappa value had decreased to 0.28 between the local orthopedic surgeons (O) and SS . TT and SS agreed in 73% of the cases (kappa statistics could not be computed because TT had no hips in groups 1 and 2).

When the Catterall grouping was reduced to 2 groups, by combining Catterall 1 and 2, and Catterall 3 and 4, the kappa values increased to 0.55 between O and SS (agreement 85%), and to 0.62 (agreement 88%) between TT and SS.

Salter-Thompson classification (Table 2)

Moderate to good agreement was obtained at the time of diagnosis, with kappa values ranging from 0.54 (O/SS) to 0.63 (TT/SS). At the 1-year follow-up we noted very low kappa values (0.18 and 0.29) between observers because there were only a few hips in group A, but the percentage agreement was good (88% and 91%). The subchondral fracture line was present in 30% of the initial radiographs. The agreement was lower between O and SS (kappa 0.45) and between TT and SS (kappa 0.59) when a fracture was present than when it was absent (kappa 0.60 for O/SS, kappa 0.65 for TT/SS).

Table 3. Inter-observer agreement, lateral pillar classification

		Agreement					Kappa$_{(W)}$
		Lateral pillar groups					
	N_1	A	B	C	N_2	%	
Primary							
O/SS	155	11	87	22	120	77	0.56
TT/SS	72	2	55	8	65	90	0.70
Follow-up							
O/SS	100	2	33	36	71	71	0.50
TT/SS	63	1	29	21	51	81	0.64

For symbols, see Table 1.

Lateral pillar classification (Table 3)

The level of agreement was moderate between O and SS when this classification was used (kappa 0.56). The agreement was good when we compared the 2 more experienced observers (kappa 0.70). This agreement was maintained at the 1-year follow-up, with kappa 0.50 (O/SS) and kappa 0.64 (TT/SS). On the initial radiographic evaluation, agreement was better between O and SS in the fragmentation phase (82%) than the initial phase (72%). Kappa could not be computed since O had no hips in group 1 in the fragmentation phase and TT had no hips in group 3 in the initial phase.

Femoral head coverage (Table 4)

We found good agreement in the assessment of primary radiographs of the Perthes' and normal hips between O/SS (ICC 0.91, 0.90) and between TT/SS (ICC 0.95, 0.86). At the follow-up, agreement was somewhat better between TT and SS concerning both the Perthes' and normal hips (ICC 0.90, 0.87) than between O and SS (ICC 0.79, 0.65). The mean inter-observer differences were low (0–2.3%) in the interpretations of the primary and follow-up radiographs and of the Perthes' and normal hips. The ranges, which include 95% of the inter-observer differences, were narrower between TT and SS than between O and SS, indicating a slightly better concordance between TT and SS (Figure 3). We found negligible differences in inter–observer agreement between the radiographic phases of the disease.

The mean -2SD femoral head coverage in normal hips was 82%, when measured by SS, 81% by O, and 80% by TT. Thus, an FHC of 80% can be regarded as a reasonable lower limit of normal variation.

CE-angle (Table 5)

We found poor agreement between O and SS in the measurement of this angle in Perthes' and normal hips at the time of diagnosis (ICC 0.51 and 0.45) and at follow-up (ICC 0.54 and 0.25), but better agreement between TT and SS at the

Table 4. Inter-observer variation in interpretation of the femoral head coverage (FHC, expressed as % coverage)

	Perthes' hip					Normal hip				
	N	ICC	Inter-obs. difference Mean	SD	Range	N	ICC	Inter–obs. difference Mean	SD	Range
O/SS										
P	139	0.91	–0.5	4.6	–9.7–8.7	136	0.90	0.0	3.3	–6.6–6.6
F–up	67	0.79	–0.5	7.2	–14.9–13.9	63	0.65	–0.7	5.3	–11.3–9.9
TT/SS										
P	71	0.95	–1.0	3.7	–8.4–6.4	71	0.86	–0.5	4.0	–8.5–7.5
F–up	57	0.90	–1.0	5.0	–11.0–9.0	52	0.87	–2.3	3.2	–8.7–4.1

O local orthopaedic surgeons, SS and TT pediatric orthopaedic surgeons. P radiographs taken at time of diagnosis and F–upradiographs taken at 1-year follow-up. ICC intraclass correlation coefficient, SD standard deviation, Inter-obs. difference inter-observer differences between O and SS (O/SS), and between TT and SS (TT/SS), Range interval which includes 95% of the inter–observer differences (mean ± 2SD), N number of patients.

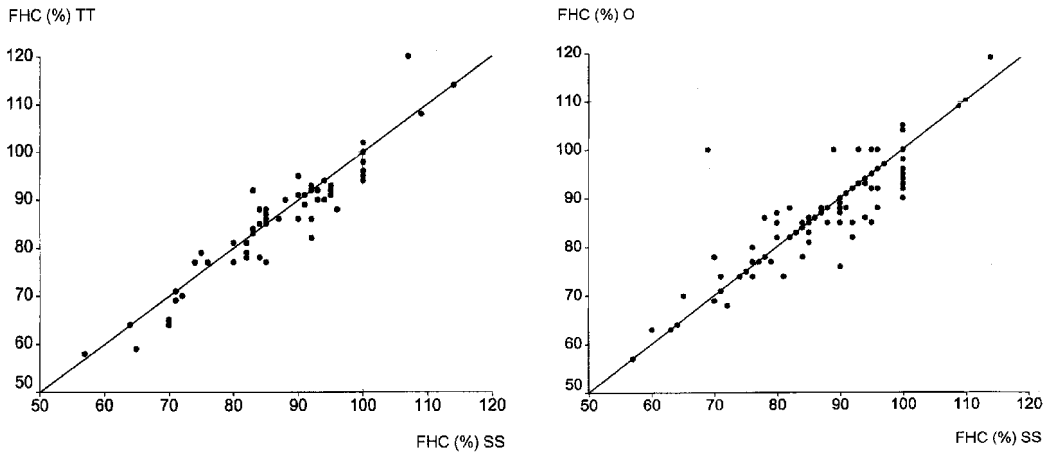

Figure 3. Scatterplot of FHC, TT versus SS and O versus SS. The line indicates perfect agreement.

time of diagnosis (ICC 0.88 and 0.72) and at the follow-up (ICC 0.78 and 0.67). The mean inter–observer differences in CE–angles were low (–2.5–2.8°), and the ranges of the differences were considerably wider between O and SS (SD 6.6–8.5) than between TT and SS (SD 2.8–3.9). These differences between the 2 pairs of observers are also clearly seen in Figure 4 and indicates poor accuracy of the CE-angles among the local orthopaedic surgeons as a group. Agreement was poorer between O and SS when the angle was measured in the fragmentation phase (ICC 0.47) than in the initial phase (ICC 0.77), but the TT/SS comparisons maintained acceptable agreement in both phases (ICC 0.78, 0.92). We found a closer agreement between the pediatric orthopaedic surgeons (TT and SS) when measuring the CE-angle in patients over the age of 5 years (ICC 0.92) than below the age of 5 (ICC 0.74).

Articulo–trochanteric distance (ATD) (Table 6)

The agreement was good in all comparisons. ICC varied little among the pairs of examiners, the time of the radiographs, and affected versus unaffected hips. The wider range of the O/SS differences (SD 2.9–3.5) than the TT/SS differences (SD 1.4–3.1) indicated a somewhat better concordance when ATD was measured by experienced observers.

Discussion

A good classification system should be easy to use, widely accepted, have good inter- and intra-rater

Table 5. Inter-observer variation in interpretation of the center-edge angle (CE-angle, expressed as degrees)

	Perthes' hip					Normal hip				
	N	ICC	Inter-obs. difference Mean	SD	Range	N	ICC	Inter-obs. difference Mean	SD	Range
O/SS										
P	82	0.51	2.8	8.5	–14.7–20.3	84	0.45	1.6	7.8	–14.0–17.2
F–up	44	0.54	–2.5	8.1	–18.7–13.7	45	0.25	–2.0	6.6	–15.2–11.2
TT/SS										
P	50	0.88	0.5	2.8	–5.1–6.1	69	0.72	1.1	3.9	–6.7–8.9
F–up	24	0.78	1.1	3.9	–6.7–8.9	49	0.67	0.7	3.8	–6.9–8.3

For symbols, see Table 4.

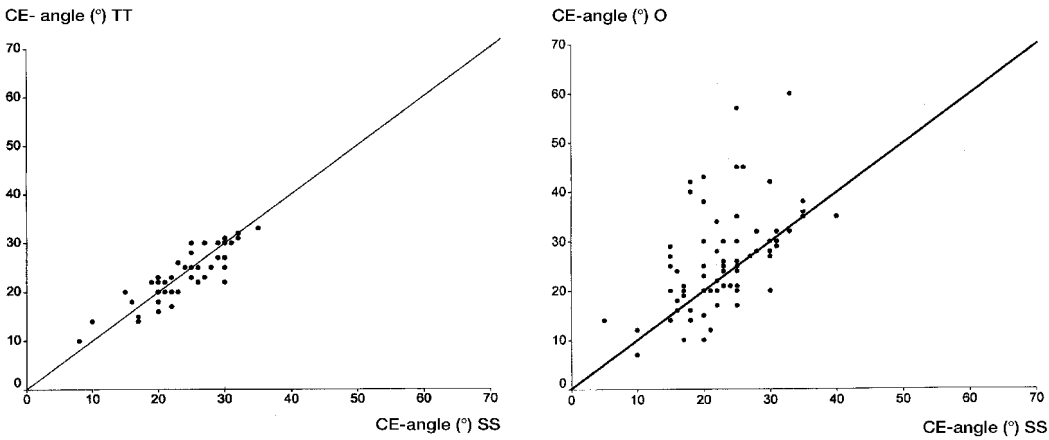

Figure 4. Scatterplot of CE-angle, TT vs SS and O vs SS. The line indicates perfect agreement.

reliability, and include the prognosis. Although the Catterall classification has been reported to have prognostic significance and is widely used, several authors have found relatively poor inter-observer agreement with this system (Christensen et al. 1978, Hardcastle et al. 1980). In our experience, this 4-group classification has not been particularly easy to use. We obtained kappa values ranging from 0.49 to 0.62 at the time of diagnosis, and agreement between experienced examiners was better than with less experienced ones (moderate agreement). This underlines the importance of pediatric orthopedic experience when assessing and treating patients with Perthes' disease. The relatively low kappa values among the less experienced observers reduce the value of Catterall classification, and we do not recommend its general use. When combining groups 1–2 and 3–4, the kappa values increased (0.55–0.62) and the level of agreement was maintained at the 1-year follow-up. It should be noted that kappa values become higher when comparing classifications with fewer categories (Altmann 1999).

Salter and Thompson (1984) presented a classification which is a simplification of the Catterall method. It is based on the extent of the subchondral fracture line and the presence or absence of an intact viable margin in the capital femoral epiphysis. This classification can be used in the early stages of the disease, when the subchondral fracture is still visible, and during the ensuing resorptive stage.

To our knowledge there has been only one report on the inter-observer reliability of the Salter-Thompson classification. Simmons et al. (1990)

Table 6. Inter-observer variation in interpretation of the articulo-trochanteric distance (ATD, in mm)

	Perthes' hip					Normal hip				
	N	ICC	Inter-obs. difference Mean	SD	Range	N	ICC	Inter-obs. difference Mean	SD	Range
O/SS										
P	90	0.81	0.3	2.9	–5.5–6.9	89	0.78	0.3	3.0	5.7–6.3
F-up	40	0.80	–0.8	3.5	–7.8–6.1	39	0.81	0.2	3.0	5.8–6.2
TT/SS										
P	47	0.88	–0.2	1.9	–4.0–3.6	49	0.92	0.0	1.4	–2.8–2.8
F-up	46	0.85	–1.3	3.1	–7.5–4.9	43	0.81	0.4	2.7	5.0–5.8

For symbols, see Table 4.

found better inter-observer agreement with this classification (kappa 0.49–0.99, agreement 68%–93%) than with the Catterall one (kappa 0.49–0.64, agreement 68%–71%). They pointed out that it was easier to use at an early stage of the disease and had a higher degree of reproducibility among more experienced observers. Our results accord with those of Simpmons et al. (1990), as there was moderate agreement (kappa 0.54, agreement 86%) between O and SS, and good agreement between the experienced examiners (kappa 0.63, agreement 89%) at the time of diagnosis. The kappa values decreased at the 1-year follow-up, although the percentage agreement was similar. In this case, the kappa values were misleading because there were only a few hips in Salter-Thompson group A.

One of the assumed advantages of the Salter-Thompson classification is its ability to predict the final extent of femoral head involvement, on the basis of the subchondral fracture line early in the disease (Salter and Thompson 1984), and thus provide the orthopedic surgeon with an accurate method for determining prognosis and treatment. This theoretical advantage seems to be of less practical value because a subchondral fracture was present in only 30% of the radiographs in our study. Moreover, we found less inter-observer agreement in hips with a subchondral fracture than without one. This may indicate that the classification is not particularly easy to use even if a subchondral fracture is present. Nevertheless, from our results and the experience of others, we think that the Salter-Thompson classification or a 2-group modification of the Catterall classification is an appropriate way to classify femoral head involvement, especially when used by experienced examiners.

Herring et al. (1992) introduced the lateral pillar classification and reported 78% agreement among observers (kappa 0.52). The inter-observer agreement was better with the lateral pillar classification than the Catterall classification. They pointed out its simplicity and found a significant correlation with the outcome reported by Stulberg et al. (1981). Ritterbush et al. (1993) also found a significantly better inter-observer reliability with the lateral pillar classification than with the Catterall classification. Farsetti et al. (1995) noted that the Herring classification was reliable and easy to use; they reported 88% agreement among observers. Indeed, the inter-observer agreement increased to 100% when the observers used a ruler to measure the height of the pillar. In a study by Podeszwa et al. (2000), moderate to good agreement among 5 observers was found (average kappa 0.51 ± 0.13) with no significant variation between observers with long experience and those with less pediatric orthopedic experience.

In our study, the lateral pillar classification had moderate (O/SS) to good (TT/SS) agreement with kappa values, ranging from 0.50 to 0.70. The highest kappa value was obtained by the more experienced observers as opposed to the findings of Podeszwa et al. (2000). These results, combined with the relative simplicity of the lateral pillar classification and its reported ability to predict outcome, indicate that this classification is useful in routine clinical work.

The most striking finding regarding the radiographic measurements was the low level of agreement between the local orthopaedic surgeons and the more experienced examiner (SS) when measuring the CE-angle. The standard deviations of the differences between O and SS were nearly 3 times larger than those between the pediatric orthopedic surgeons, indicating poor accuracy of the CE measurement among the local orthopedic surgeons. The CE-angle measurements of the normal hips showed about the same low inter-observer agreement as the Perthes' hips.

Few studies have been done on inter-observer agreement in measuring the CE-angle. Broughton et al. (1989) found that the CE-angle was helpful in assessing the dysplastic hip in children over the age of 5 years, but not in those under the age of 5, because of the ill-defined center of the femoral head. We found better agreement between the more experienced examiners when measuring the CE-angle in patients over the age of 5 than under it. Our results indicate that the CE-angle is an appropriate measurement for assessing femoral head coverage, but only when measured by examiners with adequate experience.

Severin (1941) reported that the CE-angle is abnormal when it is less than 15° in children under 13 years of age. This is in agreement with our results, since the lower limit of normal variation (mean -2SD) in normal hips, was 16° as measured by SS.

We found relatively small differences among the observers when measuring the femoral head coverage at the time of diagnosis and at the 1-year follow-up. The range of the inter-observer differences was small, indicating that the accuracy of the measurements was satisfactory, although slightly less among the local orthopedic surgeons than between the more experienced observers. Thus, among orthopedic surgeons of varying experience, FHC is a more reliable measurement for quantifying the coverage of the femoral head than the CE-angle.

The lower normal limit of femoral head coverage is an important parameter for distinguishing between normal and abnormal hips. Our results show that the lower limit of FHC (mean -2SD) is 80%, but Heyman and Herndon (195o) reported 70% in normal hips.

There was a high level of inter-observer agreement when measuring the ATD. The difference in agreement between experienced and less experienced observers was small, indicating that ATD is a reliable parameter.

No funds have been received to support this study.

Altmann D B. Practical statistics for medical research. Chapman & Hall, London 1999.

Broughton N S, Brougham D I, Cole W G, Menelaus M B. Reliability of radiological measurements in the assessment of the child's hip. J Bone Joint Surg (Br) 1989; 71: 6-8.

Catterall A. The natural history of Legg-Calvé-Perthes Disease. J Bone Joint Surg (Br) 1971; 53: 37-53.

Christensen F, Søballe K, Ejsted R, Luxhøj T. The Catterall classification of Perthes: an assessment of reliability. J Bone Joint Surg (Br) 1978; 60: 614-5.

Edgren W. Coxa plana. Acta Ortop Scand (Suppl 84) 1965.

Farsetti P, Tudisco C, Caterini R, Potenza V, Ippolito E. The Herring lateral pillar classification for prognosis in Perthes disease. J Bone Joint Surg (Br) 1995; 77: 739-42.

Hardcastle P H, Ross R, Hamalainen M, Mata A. The Catterall grouping of Perthes' disease: an assessment of observer error and prognosis using the Catterall classification. J Bone Joint Surg (Br) 1980; 62: 428-31.

Herring J A, Neustadt J B, Williams J J, Early J S, Browne R H. The lateral pillar classification of Legg-Calvé-Perthes disease. J Pediatr Orthop 1992; 12: 143–50.

Heyman C H, Herndon C H. Legg-Perthes disease. J Bone Joint Surg (Am) 1950; 32: 767-78.

McGraw K O, Wong S P. Forming inferences about some intraclass correlation coefficients. Psychol Meth 1996; 1: 30-46.

Müller M E. Die hüftnahen Femurosteotomien unter Berüchsichtigung der Form, Funktion und Beanssprüchung des Hüftgelenkes. 2 Aufl. Georg Thieme Verlag, Stuttgart 1971.

Podeszwa D A, Stanitski C L, Stanitski D F, Woo R, Mendelow M J. The effect of pediatric orthopaedic experience on interobserver and intraobserver reliability of the Herring lateral pillar classification of Perthes disease. J Pediatr Orthop 2000; 20: 562-4.

Ritterbush J F, Shantharam S S, Gelinas C. Comparison of lateral pillar and Catterall classification of Legg-Calvé-Perthes disease. J Pediatr Orthop 1993; 13: 200-2.

Salter R B, Thompson G H. Legg-Calvé-Perthes Disease. The prognostic significance of the subchondral fracture and a two-group classification of the femoral head involvement. J Bone Joint Surg (Am) 1984; 66: 479-89.

Severin E. Contribution to the knowledge of congenital dislocation of the hip joint. Acta Chir Scand (Suppl 63) 1941: 1-142.

Simmons E D, Graham H K, Szalai J P. Interobserver variability in grading Perthes disease. J Bone Joint Surg (Br) 1990; 72: 202-4.

Stulberg D, Cooperman D R, Wallenstein R. The natural history of Legg-Calvé-Perthes disease. J Bone Joint Surg (Am) 1981; 63:1095-108.

Wiberg G. Studies of dysplastic acetabula and congenital subluxation of the hip joint. With special reference to the complication of osteoarthritis. Acta Chir Scand (Suppl 58) 1939: 7-38.