Heterotopic ossification is the formation of mature lamellar bone in nonosseous tissues. It is a common complication after total hip arthroplasty (THA), occurring in 15% to 90% of cases.1 There are 24 different radiologic classification systems for heterotopic ossification, such as the Brooker, Arcq, DeLee, and Hamblen systems. The Brooker scale (Figure 1) is used in 47% of published research studies.2,3 However, the Brooker scale does not have adequate intra- or interobserver reliability.4 The amount of heterotopic ossification that interferes with hip range of motion, classified as Brooker grades III and IV, varies from 7% to 63%.5-7 Some
authors argue that small amounts of heterotopic ossification lead to significant restriction of hip mobility, whereas most believe that only a significant amount of heterotopic ossification interferes with hip mobility. In an attempt to reduce inconsistencies in grading and increase the correlation of radiographic appearance with clinical significance, Della Valle et al proposed a simplified classification system for heterotopic ossification after THA to increase intraobserver reliability and interobserver agreement (Figure 2). However, the reliability of the Della Valle classification has not been independently evaluated. Toom et al attempted to evaluate the interobserver agreement of the Della Valle, Arcq, Broker, and DeLee classification systems, but their study was grossly underpowered. The current retrospective clinical study compared the reliability of the classic Broker scale and the Della Valle classification. The Della Valle classification system consists of 3 grades, whereas the Broker scale consists of 4 grades. The Della Valle classification system provides important information on the size of islands of heterotopic ossification, whereas the Broker scale does not. The authors hypothesize that the Della Valle classification system, which is simpler (ie, fewer grades) and uses more specific descriptions (eg, size specifications), should have higher inter- and intraobserver reliability than the Broker scale when measuring heterotopic ossification after THA.

**Materials and Methods**

After institutional review board approval was obtained, the charts of 236 patients with documented radiographic heterotopic ossification at least 2 months after THA were reviewed by 3 clinicians (G.I.V, D.F.A., L.P.-S.) with experience in independent evaluation of electronic radiographs. Each reviewer noted the grade of heterotopic ossification on anteroposterior pelvic radiographs according to the Broker scale (Figure 1) and the Della Valle classification system (Figure 2). Each reviewer repeated the grading a second time at least 2 weeks after the previous review. Reviewers were blinded to their previous grades.

**Intraobserver Reliability and Interobserver Agreement**

Intraobserver reliability and interobserver agreement for each classification system were assessed with Cohen’s kappa (κ) coefficient of agreement (SAS version 9.3; SAS Institute Inc, Cary, North Carolina). Required sample size calculations were based on previously established incidences of heterotopic ossification within the THA population and an acceptable value of κ≥0.5. Inclusion of a total of 236 patients with heterotopic ossification after THA was necessary to achieve a 95% confidence interval of ±0.10 when κ≥0.5. Intraobserver reliability was assessed by comparing observations made by the same observer during the first and second evaluations. Interobserver agreement was assessed by comparing the observations made by each observer during the first evaluation. Interpretation of κ was based on the criteria of Landis and Koch, as follows: almost perfect (κ≥0.80), substantial (0.60≤κ<0.80), moderate (0.40≤κ<0.60), fair (0.20≤κ<0.40), and poor (κ<0.20).

**Results**

The Broker scale showed moderate to substantial intraobserver reliability, ranging from 0.43 to 0.71, and the Della Valle classification system showed substantial intraobserver reliability, ranging from 0.65 to 0.77 (Figure 3). No statistically significant difference was found in intraobserver reliability for either classification system for each observer, based on overlapping 95% confidence intervals.
Overall intraobserver reliability for each classification system cannot be calculated because calculations must remain independent for each observer. However, the Della Valle classification system appears to have much less variability in intraobserver agreement between observers (Figure 3).

Both classification systems showed moderate interobserver agreement. For the Brooker scale, interobserver agreement was moderate ($\kappa=0.41$; 95% confidence interval, 0.37-0.45; $P<.0001$) (Figure 4). None of the Brooker grades showed a statistically significant difference, based on overlapping 95% confidence intervals. Interobserver agreement of the Della Valle classification system was moderate ($\kappa=0.53$; 95% confidence interval, 0.47-0.59; $P<.0001$) (Figure 5). Overall, no statistically significant difference in interobserver agreement was found between Brooker grades and the Della Valle classification system, based on overlapping 95% confidence intervals. However, the best interobserver agreement for any grade was seen with grade C of the Della Valle classification system, which showed substantial interobserver reliability, based on nonoverlapping 95% confidence intervals (Figure 5).

**Discussion**

This study assessed the intraobserver reliability and interobserver agreement of 2 commonly used classification systems for radiographic evaluation of heterotopic ossification. Neither system was found to be superior, and both systems had only moderate interobserver agreement. However, the Della Valle classification system had more consistent intraobserver reliability among the 3 observers.

Both the Brooker scale and the Della Valle classifications system showed substantial reported intraobserver reliability. The authors found that the intraobserver agreement for the Brooker scale ranged from 0.49 to 0.71, depending on the observer. They found that the interobserver agreement for the Della Valle classification system ranged from 0.66 to 0.77, depending on the observer. A possible explanation for the discrepancy, especially among observers, is that both classification systems may be confusing to some observers (Figure 6). Specifically, 2 types of heterotopic ossification are not adequately described by any Della Valle grade. A small femoral or pelvic spur is...
classified as neither grade A nor grade B (Figure 7). Similarly, a large bone island that leaves little space between the island and the femur and pelvis is classified as neither grade B nor grade C (Figure 8). A similar argument can be made for the Brooker scale. A large bone island is classified as neither grade I nor grade IV (Figure 8). Hence, although pictographically the grades appear clear, in practical application, the grades are subjective, leading to variability in intraobserver reliability.

Both the Brooker scale and the Della Valle classification system showed moderate reported interobserver agreement. For the Brooker scale, $\kappa$ was 0.31 to 0.54, with mean $\kappa$ of 0.51. For the Della Valle classification system, $\kappa$ was 0.38 to 0.69, with mean $\kappa$ of 0.53. With the Della Valle classification system, evaluations by the same observer were more consistent, but overall intraobserver agreement did not change. The current data independently corroborated this observation.

**Conclusion**

For each classification system, the most significant grade (ie, Della Valle grade C and Brooker grade IV) had the best interobserver agreement. Della Valle grade C was statistically more reliable between observers, suggesting that this system may be more reliable than the Brooker scale in detecting high grades of heterotopic ossification. High grades of heterotopic ossification are easier to distinguish and may offer the most clinically relevant classifications. However, this is not universally true because a large bone island, classified as
grade I on the Brooker scale, may have the same high clinical effect as a grade IV heterotopic ossification (Figure 8). However, an ideal classification system should have high interobserver reliability, independent of the grade. Orthopedic surgeons should consider switching from the prevalent Brooker scale to the modestly more reliable Della Valle classification system when evaluating heterotopic ossification.

REFERENCES