Inter-observer and Intra-observer Reliability of the Harris Hip Scoring System

Anne Kathleen B. Ganal-Antonio and Gregorio Marcelo S. Azores

Arthroplasty Service, Department of Orthopedics, College of Medicine and Philippine General Hospital, University of the Philippines Manila

ABSTRACT

Objective. The Harris hip score (HHS) is a 100-point scale for rating pain, function, absence of deformity, and range of motion. The purpose of this study is to assess the inter-observer and intra-observer reliability of the Harris hip score among senior orthopedic residents at the Philippine General Hospital.

Methods. Twenty-four hips from 20 patients were evaluated using the Harris hip score by four senior residents from the Department of Orthopedics, Philippine General Hospital. All patients were interviewed twice in the clinic and the reliability of the HHS was evaluated.

Results. The inter-observer coefficient of concordance (Kendall coefficient of concordance W) was 0.9 for both groups of observers. The intra-observer coefficients of concordance were 0.8, 1.0, 0.9, and 0.9, for the four observers. A 0 value indicates no concordance among a set of raters while a score of 1 indicates perfect concordance. Obtaining a score greater than 0.75 represents excellent level of agreement.

Conclusion. We conclude that the Harris hip score has high inter-observer and inter-observer reliability among senior Orthopedic residents at the Philippine General Hospital.

Key Words: Harris hip scores, inter-observer and intra-observer reliability

Introduction

There is an increase in emphasis on reliable and valid measures to assess the effectiveness of treatment and rehabilitation with regard to function. A measurement is deemed reliable if the results obtained are consistent when the same entity is measured again.[1] Validity, or accuracy, is defined as the extent to which an outcome measure assesses what it claims to measure. Reliability is the basic requirement of all scientific measurement. Reliability is necessary but not a sufficient condition for validity.[2] Therefore, a measurement that is completely unreliable cannot be valid.[1]

Different hip rating scales have been developed to quantify patients' complaints. However, there is no scale that has been demonstrated to be superior to another. Most of the scales currently used have not been thoroughly tested for reliability and validity.[2] One of the problems with using a questionnaire is that it is difficult to obtain true gold standards for comparison.[3]

The Harris hip score is one of the widely accepted tools in assessing hip pain. This system was designed to evaluate different hip problems and assess different methods of treatment. It takes into account several factors, including pain, function, absence of deformity, and range of motion, in an effort to incorporate important variables into a single figure that is reproducible and reasonably objective. Pain and functional capacity, the two basic considerations, constitute the indications for surgery in patients with hip problems, thus have the heaviest weight.[4] Each category is given a corresponding score, with a perfect score of 100. Scores range from 0 (maximum disability) to 100 (no disability). The reliability and validity of the Harris hip score are unknown.[2]

In Table 1, the distribution of maximum scores in the Harris hips scores is shown. Gait is assessed based on the support necessary to walk a certain distance and the appearance of the gait after walking this distance. Motion is important because it affects function, but has little weight because active motion is more significant. Any significant deformity eliminates the four points designated for absence of deformity. A score of 90 to 100 is excellent, 80-90 good, 70-80 fair, and below 70 poor.[4]

Table 1. Distribution of maximum scores of the Harris hip score

Pain	44
Function	47
a. Daily Activities	(14)
b. Gait	(33)
Range of Motion	5
Absence of Deformity	4
Total	100

Obtaining an accurate score describing the patient's condition is crucial in managing the patient with hip pain. A

Presented at the 23rd Annual Orthopedic Residents' Research Forum, December 2005, Philippine General Hospital and at the 57th Philippine Orthopaedic Association Annual Convention, November 20-24, 2006, EDSA Shangri-La Hotel, Ortigas Center, Mandaluyong City

Corresponding author: Gregorio Marcelo S. Azores, MD
Arthroplasty Service, Department of Orthopedics
Philippine General Hospital
University of the Philippines Manila
Taft Avenue, Ermita, Manila 1000 Philippines
Telephone: +632 5548466
Email: giazores@yahoo.com

change in the scores before and after management, reflective of the patient's status, indicates the effectiveness of the treatment.[2] The purpose of the current study is to determine the inter-tester and intra-tester reliability for senior residents using the Harris hip score.

Methods

Sequential patients with hip pain seen at the Arthroplasty Clinic from August to November 2005 were considered for inclusion in the study. Clinic days were alternately assigned as inter-tester or intra-tester days. Patients who were already postoperative, had an acute fracture, or with infection at the site of the total hip replacement were excluded. A total of 24 hips from 20 patients were included.

All evaluations were conducted on their day of consult. All patients signed an informed consent and were questioned regarding demographic information, diagnosis, medications taken, and duration of pain. Feedback was also asked at the end of the study. Four senior residents conducted the Harris hip score for the duration of the study. Prior to the conduction of the test, the observers were oriented on the proper administration of the Harris hip score.

Figure 1 shows the data collection form used in the study.

As the patient came in for consultation, he was alternately assigned to different observers. There were two observers during each clinic day. For intra-observer data collection, the resident evaluated the same patient twice, with time lapse between two assessments. For inter-observer data collection, the resident evaluated the patient once and was assessed a second time by another resident, with time lapse between measurements. New evaluation forms were given to the residents every time they assessed a patient. The decision to conduct the interview on the same day rather than having a longer interval between observations was due to concern for potential changes in the clinical status of the patients during the time interval.

Total Harris hip scores were analyzed using Kendall's W (Kendall coefficient of concordance W). Range of values of a Kendall's W is from 0-1. A 0 value indicates no concordance among a set of raters while a score of 1 indicates perfect concordance. There were two sets of computations for the inter-observer part, one set for raters 1 and 2 and another for raters 3 and 4. For the intra-observer concordance, there were four computations having each rater rating four patients at a time. A correlation coefficient less than 0.4 was considered poor, between 0.4 to 0.75, fair, and greater than 0.75, excellent.[1,5]

Statistical analysis was performed for the total scores of the Harris hip scores only and not the categorical scores (excellent, good, fair, and poor). Although Harris hip scores can be converted into descriptive terms, this process can diminish the reliability of the hip scores.[6] Bach investigated the influence of descriptive and numeric outcomes for inter-observer and inter-score correlation of five different hip scores. Higher inter-observer reliability and inter-score correlation was found when compared with the category system.[6] Only the total scores of the Harris hip score were statistically analyzed.

Results

Twenty patients with 24 symptomatic hips were seen and enrolled in the study between August and November 2005. Sixteen hips were in the intra-observer group and 8 hips were assessed in the inter-observer group. The mean age of the patients included was 42.3 years (range, 19 to 73 years). Table 2 shows the sex distribution, having six patients (30 per cent) being men, while the rest were females.

Table 2. Sex distribution

Sex	Percentage
Male	30% (6)
Female	70% (14)
	100% (20)

Of the 20 enrolled patients, 8 (40 %) had idiopathic osteoarthroses, 3 (15 %) had avascular necrosis of the hip, 5 (25%) had osteoarthrosis due to childhood hip disease, 2 (10%) rheumatoid arthritis, and 5 (25 %) had other diagnoses. Table 3 shows the distribution of level of activity of the patients. Eighteen (90%) were community ambulators, while 2 (10%) were household ambulators. There were no bed-ridden patients in the study.

Table 3. Distribution based on level of activity

Activity Level	Percentage
Bed-ridden	0 (0)
Household ambulator	10% (2)
Community ambulator	90% (18)
	100% (20)

Table 4 shows that four (20%) were smokers. Table 5 indicates that four (20%) were drinkers and the rest did not consume alcohol. Sixteen (80%) were unemployed.

Table 4. Smoker vs. Non-smoker

	Percentage
Smoker	20% (4)
Non-smoker	80% (16)
	100% (20)

Table 5. Drinker vs. Non-drinker

	Percentage
Drinker	20% (4)
Non-drinker	80% (16)
	100% (20)

Figure 1. The Data Collection Form

Inter-observer and intra-observer reliability of the Harris Hip Scoring system in patients of the PGH-Arthroplasty clinic

Name:				**Px #**	**Hip: R L B**
Age: years	**Sex: F M**	**CN:**		**XN:**	
Wt: kgs	**Ht: cm**	**Observer: 1 2 3 4 5 6 7 8 9 10**		**Occupation:**	
Activity Level: **[ ] Bed ridden** **[ ] Household ambulatory** **[ ] Community ambulator**		**Smoker**	**Y** **N**	**[ ] Unemployed** **[ ] Heavy laborer** **[ ] Clerical** **[ ] Specify:**	
		Drinker	**Y** **N**		
		Medications: **[ ] Steroids [ ] NSAIDs** **[ ] Specify:**			
Menopause: Y N **When: years ago**		**Number of pregnancies:** **1 2 3 4 [ ] others:**		**Duration of Hip Pain:** **[ ] < 1 month [ ] 1-6 months** **[ ] 6-12 months** **[ ] 1-5 years [ ] >6 years**	
Assessment:					

Date: **Intra Inter**
Patient number: **Observer Number:**

Pain (44 points)		
None/ignores	***44***	
Slight, occasional, no compromise in activity	***40***	
Mild pain, no effect on average activities, rarely moderate pain with unusual activity, may take aspirin	***30***	
Moderate pain, tolerable but makes concessions to pain. Some limitations of ordinary activity or work. May require occasional pain medication stronger than aspirin	***20***	
Marked pain, serious limitations	***10***	
Totally disabled, crippled, pain in bed, bed-ridden	***0***	
Function (33 points)		
Gait (Walk maximum distance)		
Limp: None	***11***	
Slight	***8***	
Moderate	***5***	
Severe or unable to walk	***0***	
Support: None	***11***	
Cane for long walks	***7***	
Cane most of the time	***5***	
1 Crutch	***4***	
2 canes	***2***	
2 crutches or unable to walk	***0***	
Distance walked: Unlimited	***11***	
6 blocks	***8***	
2-3 blocks	***5***	
Bed and chair	***0***	
Functional Activities (14 points)		
Stairs: Normally without using a railing	***4***	
Normally using a railing	***2***	
In any manner	***1***	
Unable to do stairs	***0***	
Socks/ tie shoes: With ease	***4***	
With difficulty	***2***	
Unable	***0***	
Sitting: Comfortably on ordinary chair for one hour	***5***	
On a high chair for half an hour	***3***	
Unable to sit comfortably in any chair	***0***	
Enter public transportation	***1***	
Not able to use public transportation	***0***	
Flexion		
Extension		
Abduction		
Adduction		
Internal rotation		
External rotation		
LLD		
Total		

Table 6 shows the distribution based on medication use. Among the 20, 2 (10%) were taking steroids, while 9 (45%) were on NSAIDs, and the rest did not take any medication. One (5%) had symptoms less than 1 month, 6 (30%) had symptoms between 1-6 months, 4 (20%) with symptoms from 6 to 12 months, 6 (30%) with symptoms for 1 to 5 years, and 3 (15%) had symptoms for more than 6 years.

Table 6. Distribution based on medication use

Medication	Percentage
None	35% (7)
Steroids	10% (2)
NSAIDs	45% (9)
Others	10% (2)
	100% (20)

Between Observers

There were two groups of four hips evaluated by two different observers at a time for inter-observer reliability. Kendall coefficient of concordance, W, was used to compute for agreeability between observers. Table 7 shows the inter-class coefficient of 0.9 concordance between raters 1 and 2 while table 8 shows an inter-class coefficient of 0.9 between raters 3 and 4.

Table 7. Inter-class Coefficient of Concordance of Raters 1 and 2 for Hips 1 to 4

	Hip 1	Hip 2	Hip 3	Hip 4	Kendall W
Rater 1	96	51	45	93	
Rater 2	100	41	44	71	
					0.9

The Harris hip score has excellent inter-observer reliability with Kendall W of 0.9 for both groups of observers. This means that there is excellent agreement among the observers.

Within Observers

There were four computations for intra-observer group, one for each rater evaluating four patients each. Again, the Kendall coefficient of concordance, W, was used to compute for agreeability between observers. Tables 8 to 11 show the results for the four observers.

The intra-observer concordance, W, shows excellent agreement between two observations for all observers.

Table 8. Inter-class Coefficient of Concordance of Raters 3 and 4 for Hips 5 to 8

	Hip 5	Hip 6	Hip 7	Hip 8	Kendall W
Rater 3	86	46	68	62	
Rater 4	86	48	95	65	
					0.9

Table 9. Intra-class Coefficient of Concordance for Rater 2 for Hips 5 to 8

Rater 2	Hip 5	Hip 6	Hip 7	Hip 8	Kendall W
1	30	94	50	100	
2	26	96	81	100	
					1.0

Table 10. Intra-class Coefficient of Concordance for Rater 3 for Hips 9 to 12

Rater 3	Hip 9	Hip 10	Hip 11	Hip 12	Kendall W
1	70	57	52	37	
2	80	44	54	38	
					0.9

Table 11. Intra-class Coefficient of Concordance for Rater 4 for Hips 12 to 16

Rater 4	Hip 13	Hip 14	Hip 15	Hip 16	Kendall W
1	83	100	79	72	
2	93	100	76	82	
					0.9

Discussion

Health–measurement scales may be classified as generic health-status measures or as disease-specific measures. Disease-specific measures, such as the Harris hip score, focus on particular complaints attributable to the disease or the condition, and reflects clinical change. Generic health-status measures, such as Short Form-36 (SF-36), are intended to measure all aspects of a patient's health, including physical function, emotional, mental, and social function.[3] Disease-specific scales are more responsive than generic health-status measure for evaluation of outcomes of orthopedic procedures.[1] The Harris hip score is closely associated with SF-36, which is a widely accepted generic self-administered test of health-related quality of life. In a comparison between the Harris hip scores and SF-36, results showed higher responsiveness ratios for the Harris hip scores than generic scales like the SF-36.[7] Hoeksma compared the responsiveness between the Harris hip score, SF-36, and a test for walking speed. The Harris hip score was more responsive, more sensitive and specific, detecting a small improvement of 8% change from baseline. This suggests that the Harris hip score is a more suitable instrument in evaluating change in hip function in patients with osteoarthritis of the hip.[8]

The Harris hip score is an interviewer-administered questionnaire, allowing the interviewer to ensure that the questions are interpreted appropriately and that the patient is the actual respondent to the questionnaire. The high responsiveness of the Harris hip score is probably because it combines both observational and self reported items.[7] However, there is concern regarding the interviewer's effect on the patient's responses to the questions. Social desirability bias occurs when a respondent answers

questions with what he thinks are acceptable answers. This phenomenon is a concern in orthopedics because many patients may want to please their surgeons.[9] Being an observational assessment, the Harris hip score is less sensitive to a patient's subjective bias for providing socially desirable answers against actual recovery.[8] These questionnaires are also more expensive to administer because it is often necessary to train the interviewer.[9]

Why should measurements be reliable? A change in a reliable clinical measurement can be attributed to a true change in the clinical status of the patient. Reliable measurements can improve the efficiency of clinical trials and is a minimal requirement for valid outcomes.[1] Having reliable measurements allows clinicians to be increasingly confident in attributing a difference between two measurements to a change in the clinical status of the patient.

Unreliable measurement can arise from the patient, the procedure, and the clinician, with each factor considered separately to reduce measurement variability. Patient variability occurs because of biologic inconsistencies, which varies depending on the patient's state or condition. The decision of administering the test in one afternoon was based on the potential change in the patient's status over a prolonged interval of time. The change in the status will affect the outcome of the Harris hip score, thereby affecting the inter-observer and intra-observer results. In a study by Wright, 38% of patients indicated that their condition either slightly improved or worsened even with an interval of only 2 weeks.[9]

Inconsistencies in the procedure, the observer's performance, or the equipment used may lead to procedural variability. To decrease procedural variability, the method should be described in detail to allow replication of the technique for repeated measurements. In this study, the senior residents were already familiar with the scoring system to be evaluated; they were nonetheless oriented on the administration of the test to ensure a more consistent procedure in conducting the scoring system.

The last source of variability is the clinicians observing, extracting, and interpreting information. Reducing clinician variability is accomplished by clarifying the observational process and conversion criteria. Measurements vary less when performed under ideal circumstances. Another bias is related to skill and experience. Skills improve with practice, and more accurate measurements are obtained with repeated assessments. However, measurement of variability among highly experienced clinicians will not accurately reflect the usual variability of inexperienced clinicians.[1] For a study to be valid, results should be compared against results obtained by a comparable observer.[10] Ideally, the system can be used easily by both experienced and inexperienced observers. In this study, residents in their senior year of training were chosen as observers, all of them with comparable experience. Though they already had experience in using the system, they are not yet considered specialists, thus reflecting the usual variability of a less experienced clinician. However, their previous experience would in theory result in more precise measurements.

In general, intra-clinician variability is smaller than inter-clinician variability. Variability can come from the number of categories in a scale or from the specifications of conversion criteria. Having more categories allows more clinical distinction but may be offset by greater variability. Ambiguous, incomplete criteria may also lead to substantial variability. For example, "limp" can be graded slight, moderate, or severe but the absence of criteria for these ratings may lead to variability. Another is walking ability, which has no clear guide for categorizing each patient.[1,4] Other categories in the Harris hip score, such as wearing socks and tying shoelaces, or the concept of a "block", may not be universally applicable in the Filipino setting. This presents a problem in conducting the study and bias may be introduced when administering the system. Categories are translated into the Filipino language and, while doing so, information can either be leading or misleading. Having only four major categories in the Harris hip score, variability is decreased. However, despite having ambiguous criteria, our data suggests a high level of agreement between observers. In this study, both inter-observer and intra-observer concordance coefficients were excellent. There is little variation in both the inter- and intra-observer scores.

The main limitation of the present study was the number of patients involved in the study. Only 24 hips were evaluated, 8 in inter-observer and 16 in the intra-observer group; therefore, results cannot be generalized. The solution is to increase the number of people in each group of the clinical trial.

Another limitation of this study is observer bias. Outcomes should be evaluated by qualified independent observers, whose involvement is limited only to objective measurement of the results.

We recommend conducting a subsequent study to compare the Harris hip score with other disease-specific measures. We also recommend translating the Harris hip scoring system into Filipino or conducting cultural adaptability study for Filipinos.

Conclusion

The Harris hip score had good intra- and inter-observer reliability in this study. This scoring system provides reproducible assessment of patients with hip pain and could be used in monitoring changes in the patient's status.

References

1. Wright JG, Feinstein AR. Improving the reliability of orthopaedic measurements. J Bone Joint Surg Br. 1992; 74(2):287-91.
2. Wright JG, Young NL. The patient-specific index: asking patients what they want. J Bone Joint Surg Am. 1997; 79(7):974-83.
3. Jaglal S, Lakhani Z, Schatzker J. Reliability, validity, and responsiveness of the lower extremity measure for patients with a hip fracture. J Bone Joint Surg Am. 2000; 82-A(7):955-62.
4. Harris WH. Traumatic arthritis of the hip after dislocation and acetabular fractures: treatment by mold arthroplasty. J Bone Joint Surg Am. 1969; 51(4):737-55.
5. Ali AM, Angliss R, Fujii G, Smith DM, Benson MK. Reliability of the Severin classification in the assessment of developmental dysplasia of the hip. J Pediatr Orthop B. 2001; 10(4):293-7.
6. Bach CM, Feizelmeier H, Kaufmann G, Sununu T, Gobel G, Krismer M. Categorization diminishes the reliability of hip scores. Clin Orthop Relat Res. 2003; (411):166-73.
7. Shields RK, Enloe LJ, Evans RE, Smith KB, Steckel SD. Reliability, validity, and responsiveness of functional test in patients with total joint replacement. Phys Ther. 1995; 75(3):169-79.
8. Hoeksma HL, Ven den Ende CH, Ronday HK, Heering A, Breedveld FC. Comparison of the responsiveness of the Harris Hip Score with generic measures for hip function in osteoarthritis of the hip. Ann Rheum Dis. 2003; 62(10):935–8.
9. Wright JG, Young NL, Waddell JP. The reliability and validity of the self-reported patient-specific index for total hip arthroplasty. J Bone Joint Surg Am. 2000; 82(6):829-37.
10. Gartland JJ. Orthopaedic clinical research: deficiencies in experimental design and determinations of outcome. J Bone Joint Surg Am. 1988; 70(9):1357-64.