Inter- and intra-measurer error in the measurement of Reimers’ hip migration percentage

S. Faraj, W. G. Atherton, N. S. Stott

From Starship Children’s Hospital, Auckland, New Zealand

Reimers’ hip migration percentage is commonly used to document the extent of subluxation of the hip in children with spasticity. In this study, two measurers, with six months paediatric orthopaedic experience, measured the migration percentage on 44 pelvic radiographs of children with cerebral palsy, aged between two and eight years. Unknown to the measurers, each radiograph was duplicated, giving 22 non-identical radiographs (44 hips) which were measured twice at time 0 and twice six weeks later. The intra-measurer, intra-sessional absolute differences between the first and second measurements ranged from 0% to 23%, with median values of 2.5% to 3.6%. The intra-sessional median absolute differences were not statistically different between the two measurers and measuring sessions (p = 0.42, Kruskal-Wallis test). The inter-sessional absolute differences for measurements made by the same measurers ranged from 0% to 18% with a median absolute difference of 1.7% to 3.2%. Overall, only 5% of the intra-measurer measurement differences, within and between sessions, were above 13%. Repeated measurements by one measurer over time must, therefore, vary by more than 13% in order to be 95% confident of a true change. The inter-measurer error was higher with median absolute differences between the two measurers’ measurements of the same hip of 3.25% to 5% (0% to 26%) and a 95th upper confidence interval of 21% to 23%. Averaging the four separate measurements over the two sessions reduced the inter-measurer error to a median absolute difference of 2.8%, but did not improve the 95th upper confidence interval, which measured 22.4%.

Such inter-measurer errors may be clinically unacceptable.

Subluxation and dislocation of the hip are common in children with cerebral palsy.1-3 The effects of dislocation can include pain, contractures and walking difficulties. In 1980, Reimers3 described a hip migration percentage, which could be used to document the extent of hip subluxation lateral to the acetabulum. He proposed that this measurement was more reliable for spastic hip subluxation than the centre edge angle (CE angle), which he had found to be more position dependent and nonlinear. He estimated a standard error of ±10% for the migration percentage, based on an estimate of the errors involved in measuring a line to the nearest millimetre.

The migration percentage is widely used in the cerebral palsy literature, but the reliability of this measurement has only been formally studied, to our knowledge, by one group of authors.4 The aim of the study was to evaluate the inter-measurer and intra-measurer error of Reimers’ migration percentage when used by orthopaedic trainees who have completed six months of elective paediatric orthopaedic training.

Patients and Methods

Radiographs. We duplicated 22 pelvic radiographs (44 radiographs, 88 hips) which had been taken in children with cerebral palsy, aged between two and eight years. As the radiographs were presented in a random fashion to the two measurers, they were unaware that the radiographs had been duplicated. The duplicate radiographs were copied in reverse to lessen the chance that a measurer would recognise that the radiograph had been duplicated. The radiographs chosen had to have clear visualisation of both hip joints with both in the neutral position with respect to adduction and abduction. Radiographs taken after previous hip or pelvic surgery were excluded.

Measuring. Both measurers chosen for the study were orthopaedic trainees who had spent six months working in a paediatric orthopaedic unit in a tertiary level hospital. Prior to
the measuring session, the measurers completed a discussion on the migration percentage and reviewed the measuring technique. They then carried out a practice measurement on a sample hip radiograph. Each measurer used their own markers (sharp wax film-marker), ruler and the same goniometer. The radiographs were measured in a random sequence by the first measurer, and then the markings were removed with an alcoholic swab. The second measurer then measured all the radiographs using the same goniometer, removing all the markings at the end of the measurement session. Six weeks later, the measurements were repeated with the radiographs presented in a different order. No discussion or feedback was allowed between the two measurement sessions.

**Measurement.** Figure 1 shows an example of the measurement. Hilgenreiner’s line and Perkins’ line were drawn as defined by Reimers. The migration percentage was calculated as the part of the femoral head outside Perkins’ line divided by the total width of the femoral head multiplied by 100. The measurements were repeated after six weeks with the same previous parameters.

**Measurement reliability.** The intra-measurer reliability within each session was determined by calculating the difference between the measurer’s migration percentage score for copy 1 of the radiograph and the score for copy 2 i.e. the identical duplicated radiograph. The intra-measurer reliability between sessions was calculated by comparing the average of the duplicate measurements for each session by each measurer. The inter-measurer reliability within one session was determined by calculating the average difference between the migration percentage scores for each measurer.

**Statistical analysis.** The data were placed on an Excel spread sheet and analyzed using SPSS 11.0 software and Sigma Plot 8.0. The differences between the measurements were calculated for each session. These differences were then converted into absolute numbers and graphed as box and whisker plots using Sigma Plot 8.0 software. These plots show the median absolute difference, 25th to 75th confidence intervals (solid box) and 10th to 90th confidence intervals (whiskers). Two non-parametric tests, Kruskal-
Intra-measurer, inter-sessional differences in measurement of Reimers’ hip migration percentage. Box and whisker plots show the inter-sessional absolute differences for measurer one and measurer two. The solid box represents the 25th to 75th percentile, the whiskers represent the 10th to 90th percentile and symbols represent outlying points. The median absolute differences (50th percentiles) are displayed as horizontal lines and range from 3.2% for measurer one to 1.7% for measurer two. The highest inter-sessional error is 18% for measurer one and 13.6% for measurer two.

Results
Intra-measurer error
Figure 2 is a graphic representation of the intra-measurer, intra-sessional error for both measurers and both sessions. The absolute difference between the first and the second measurements, recorded by one measurer within one session, ranged from 0% to 23% with a median absolute difference which ranged from 2.5% (measurer two, session two) to 3.6% (measurer one, session one). The median absolute differences were not significantly different between either sessions or measurements (p = 0.42, Kruskal-Wallis non-parametric ANOVA). Of the absolute differences in measurements, 75% were within the range 0% to 6.2% for measurer one and 0% to 5% for measurer two, when the data for the two sessions were combined. Similarly, 90% of the absolute differences in measurements were within the range 0% to 9.3% for measurer one and 0% to 8% for measurer two, when the data from the two sessions were combined. Pooling of the intra-sessional absolute differences for the two measurers and the two sessions led to a median absolute difference of 3% and a 95th percentile of 12.7%.

Figure 3 shows the intra-measurer, inter-sessional absolute differences for the two measurers. The absolute differences between measurements made in session one and session two ranged from 0% to 18% for the two measurers. The inter-sessional absolute differences for measurer one had a higher median of 3.2%, compared to a median of 1.7% for measurer two, however this difference was not statistically significant (p = 0.12, Mann-Whitney U). Of the inter-sessional absolute differences in measurements, 75% were within the range 0% to 6% for measurer one and 0% to 3.4% for measurer two. The 90th percentile was similar for both measurers at 8.7% for measurer one and 7.9% for measurer two. A number of outlying points meant, however, that the 95th percentile was 12.9% for measurer one and 12.6% for measurer two.

Inter-measurer error. A systematic error was noted with measurer two recording a higher migration percentage than measurer one in 37 of 44 hips in the first session and in 30 of 44 hips in the second session. Figure 4 shows that the median absolute difference between the two measurers’ measurements of the same hip was 5.0% for session one and 3.2% for session two. This difference was not significant (p = 0.13, Mann-Whitney U). The individual absolute differences had a wider range with the highest inter-
measurer absolute difference being 26.5%. Of the differences in measurements by the two separate measurers, 90% were below 15% when data from the two sessions were combined. A small number of outlying points meant that the 95th percentile was much higher at 23% in session one and 21% in session two. Averaging the mean scores from session one and session two to give an overall average of four separate measurements, led to a reduction in the median inter-measurer absolute difference to 2.8%, but no change in the 95th percentile, which measured 22%.

Discussion
Reimers’ migration percentage is commonly used in the management of subluxation of the hip in cerebral palsy. Studies have used this percentage to monitor the effects of interventions such as selective dorsal rhizotomy, hip abduction bracing and soft-tissue surgery on the progression of subluxation. Several studies have used changes greater than 10% as significant based on an estimated standard error of ±10% suggested by Reimers in 1980.1

In our study, we have assessed the reliability of measurement of the hip migration percentage, both within and between sessions, for two measurers with six months of elective paediatric orthopaedic experience. This level of experience is what might be expected for a recently graduated orthopaedic surgeon following four years of training in Australia or New Zealand. We found that a change of greater than 13% in repeated measurements by one measurer is required in order to be 95% confident of a ‘true’ change. Eklof, Ringertz and Samuelsson,10 in a study of migration percentage normative values, found that a difference of greater than 12% between sides in a normal radiograph of the pelvis was highly unlikely. They hypothesised that one possible cause for this finding would be measurement error although this was not formally assessed in their study. Our data lends support to their hypothesis.

Parrott et al4 recently reported a study similar to ours, using five experienced measurers and 20 radiographs, which were measured twice over a period of two weeks. They reported data based around the 95th percentile, concluding that changes in the migration percentage of greater than ±8.3% can be used to represent ‘true’ change when the measurement is undertaken over time by one measurer. The issue of ‘true’ change, i.e. change that is outside the expected error range, is an important one for clinicians. This is particularly so when operative decisions are based on a single measurement. Our study documented several instances of large differences between two measurements made by one measurer within one session. The causes for these large differences are difficult to separate out, but probably represent a combination of true measurement variability and other errors such as a temporary lapse in concentration leading to an isolated incorrect measurement or incorrectly recording an accurate measurement, for example, 23 rather than 32. Such random errors are probably just as likely to occur in a busy clinical setting, where there are many distractions and pressures on time. Whatever the reason, these ‘differences’ in measurements could have been interpreted as either a significant improvement or deterioration if the 95th percentile were taken as the indicator of change.

Broughton et al11 have proposed that change in successive measurements over time is a better basis for clinical decisions than a single measurement. Our study could not test this statement. We did find, however, that the differences between measurements of the same radiograph by different measurers had significantly higher 95th percentiles of 21% to 23%. This suggests that repeated measurements should always be made by one measurer to reduce error. Averaging of repeated measurements could theoretically reduce error, however we were unable to show any reduction in inter-measurer error using this approach in the current study.

In summary, this study has shown that the measurement of migration percentage on hip radiographs can have significant errors. Our measurers had less experience in undertaking this measurement than those in the study by Parrott et al4 but would have had similar expertise to recently qualified orthopaedic surgeons. The inter-measurer errors were such that measurements of the same radiograph made by different measurers had errors that could be outside clinical acceptability. Other major issues, which we cannot address in this study, include the validity of this measurement and the relationship of both the reliability and validity of the measurement to temporal changes in femoral head ossification.12

The authors would like to thank Christine Ganly for her expert secretarial help.

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