Send Orders for Reprints to reprints@benthamscience.ae

The Open Nursing Journal, 2015, *9*, 33-41 33

Open Access

Reliable Assessors of Infant Cranial Asymmetry in Child Health Care

Freda Lennartsson[*,1], Göran Wennergren[2] and Per Nordin[3]

[1]*Department of Pediatrics, University of Gothenburg, 416 85 Gothenburg, Sweden*

[2]*Department of Pediatrics, University of Gothenburg, Queen Silvia Children's Hospital, 416 85 Gothenburg, Sweden*

[3]*The Skaraborg Institute for Research and Development, Stationsgatan 12, 541 30 Skövde, Sweden*

Abstract: *Introduction:* Acquired cranial asymmetry is prevalent in infants today and largely attributed to the supine sleep position recommended for infant safety. There is a risk of permanent cranial asymmetry, so prevention and early detection are important. A prevention project was initiated in Sweden, and an intervention was planned. The aim of this study was to evaluate reliability of assessors judging infant cranial asymmetry in order to evaluate if they could be considered reliable interchangeable assessors in the planned intervention.

Materials and Methodology: Five assessors were taught how to assess infant cranial asymmetry using illustrated severity assessments. They were intra-rater and inter-rater reliability tested by taking a photograph test-retest and an infant test. Agreement matrices were devised to illustrate assessor agreement based on both *type* and *degree* of cranial asymmetry. Agreement based on degree of asymmetry was analyzed by calculating AC2 using quadratic weights. Results were adjusted to arrive at the perceived genuine agreement and interpreted according to Landis and Koch's strength of agreement intervals.

Results: In the photograph test, mean percentage of perfect intra-rater agreement was 73. Adjusted mean intra-rater AC2 was 0.69 [0.63; 0.76], and adjusted inter-rater AC2s were 0.72 [0.64; 0.81] and 0.71 [0.63; 0.79]. In the infant test, the adjusted inter-rater AC2 was 0.73 [0.60; 0.87]. Results indicate substantial strength of assessor agreement.

Conclusion: Assessors were reliable and interchangeable. In a larger clinical context, results indicate that educating child health care nurses to assess infant cranial asymmetry can be used for early detection.

Keywords: Agreement measure, child health centers, infant cranial asymmetry, nonsynostotic plagiocephaly, nursing assessment, reliability.

INTRODUCTION

Nonsynostotic plagiocephaly (NSP) is an acquired cranial asymmetry. It develops pre- or postnatally from external pressure on the moldable skull of infants [1]. NSP falls into three main groups: plagiocephaly - skewed occipital flattening, brachycephaly - symmetric occipital flattening, and combined plagiocephaly-brachycephaly [2]. A rise in incidence of NSP was noted in several American tertiary care centers in the 1990s. This was largely attributed to parents following the recommendation to place infants supine when they sleep in order to prevent Sudden Infant Death Syndrome (SIDS) [3-5]. In a recent Canadian cohort study, it was estimated that 47% of infants had some degree of NSP at 7 to 12 weeks of age, and 22% of the cases were moderate or severe [6].

Since it is important that parents follow the infant supine sleep position recommendation for SIDS prevention, efforts to prevent NSP are needed. Early detection and early intervention are important because after six months of age there is a risk of permanent cranial asymmetry [7]. In a recent study of infants aged five to six months with moderate to severe NSP in the Netherlands, 74% of infants in the treatment group and 77% of infants in the natural course group had not fully recovered six months later [8]. In a New Zealand follow-up study of children diagnosed with NSP in infancy, 39% of the measured children were not in the normal range at three to four years of age, and 4% of them were in the severe range [9]. In an American study of youth aged twelve to seventeen, 2% had NSP, 38% of whom had abnormal facial characteristics [10].

A prevention project was initiated in the Swedish County of Skaraborg in 2008. NSP prevention guidelines for child health nurse clinicians were developed [11], tested in a pilot study [12], and revised. Then a short educational program for nurses was developed incorporating these guidelines. This is not an advanced program, but it went beyond the standard recommendations. An intervention study was planned to explore how well child health nurse clinicians exposed to this new program could prevent and detect incipient cranial asymmetry. Assessors from outside the child health clinics were recruited for the intervention study. They were trained to assess infant cranial asymmetry the same way we plan to train child health nurse clinicians. As part of our method development, we aspired to examine the assessors' reliability as precisely as possible.

*Address correspondence to this author at Spjutgatan 5, SE-533 34 Götene, Sweden; Tel: +46-511-340038; Fax: +46-500-478392;
E-mails: freda.lennartsson@gmail.com, freda.lennartsson@gu.se

1874-4346/15

Reliability concerns whether a measurement produces similar results under consistent conditions; and items such as physical findings often rely on some degree of subjective interpretation by observers [13]. Quantifying the severity of cranial asymmetry is difficult because many providers use an expert opinion approach that is not well documented and varies by practitioner [14]. Visual rating systems are available for classifying type and severity of NSP and potentially useful [15]. Argenta proposed a system that classifies plagiocephaly and brachycephaly exclusively by the clinical appearance of the child's head [16]. Hylton-Plank *et al*. introduced a 5-level severity scale [17]. Cranial Technologies developed Plagiocephaly and Brachycephaly Severity Assessments with four levels that include five and three scales respectively [18]. Feijen *et al*. found a significant correlation between the subjective assessment of cranial shape by physicians using Argenta's scale system and plagiocephalometry results [19], but the diagnosis of cranial asymmetry remains imprecise [15].

There are several recent studies where cranial asymmetry assessments were done using visual rating systems. Cavalier *et al*. trained primary care pediatricians prior to participation in a plagiocephaly study of newborns by teaching them about the diagnosis of plagiocephaly using Argenta's scale system, but did not reliability test the physicians prior to the study [20]. Mawji *et al*. dealt with reliability by having the two assessors spend eight hours or more conducting assessments together with clinicians at the Head Shape Clinic prior to the study using Argenta's scale system. The two assessors also convened halfway through the year and near the end of data collection to ensure consistency in the assessments [6]. Öhman did a reliability test where 39 physical therapists assessed 30 photographs of infants using the Severity Assessment for Plagiocephy. The photographs were assessed on two occasions with at least one week in between test occasions. Öhman did not train the physical therapists before testing them [21].

The aim of this study was to examine reliability of the assessors in order to evaluate if they could be considered reliable interchangeable assessors in a planned intervention study.

MATERIALS AND METHODOLOGY

Materials

The Severity Assessment for Plagiocephaly and the Severity Assessment for Brachycephaly, developed by Cranial Technologies and available online [18], were chosen because they are potentially very useful for the child health care setting. They consist of sets of pictures which serve as a guide because trait development is illustrated stepwise. The Severity Assessment for Plagiocephaly has five sets of pictures, and the Severity Assessment for Brachycephaly has three sets of pictures. Each set of pictures serves as a reminder of a different feature that is important to examine when assessing cranial asymmetry. Although they do not represent exact measurements, the four consecutive pictures in each set illustrate the increasing asymmetry of that particular feature. The increasing asymmetry of each feature is on a continuum, but arbitrarily assigned cut-off levels to indicate mild, moderate and severe are generally accepted [15, 17]. Scores from 0-3 representing the stepwise changes of increasing asymmetry are attached to the four pictures in each set creating a scale, where 0 designates no asymmetry, 1 designates mild asymmetry, 2 designates moderate asymmetry, and 3 designates severe asymmetry.

For guidance in the assessment, three sets of pictures could be used when assessing photographs. These three sets of pictures were copied onto the photograph test score sheets - plagiocephaly posterior flattening, brachycephaly posterior flattening, and brachycephaly lateral view flattening. Six sets of pictures were copied onto infant test score sheets.

Fifty photographs of infant heads depicting a variety of asymmetry features with differing degrees of asymmetry were selected from plagiocephaly and brachycephaly websites. Some photographs were considered to depict normal head shape. Nineteen vertex view photographs, i.e. seen from above, and 31 lateral view photographs were used. There were more lateral view photographs because these depict many features of a child that assessors might be misled by inadvertently, whereas vertex view photographs depict few features. Several photographs that were not straightforward were purposely selected to encourage assessors to deliberate very carefully. For instance, in one photograph it was unclear if the head was actually skewed or merely turned to the side.

Participants

The five assessors recruited externally for the intervention, four registered nurses and one medical secretary, participated in the reliability study. There were no particular educational prerequisites for these assessors. They were taught how to do assessments and then their reliability was tested by assessing cranial asymmetry in photographs and infants. All five assessed photographs, and four of them assessed infants. One was missing when infants were assessed due to late recruitment.

Six infants participated in the study when they attended their four month child health check–up. It was a convenience sample on a specific day when we could get a number of infants and the assessors together. Parents gave permission for their infant to participate and held their infant in the requested position during the assessments.

The project leader developed the photograph and infant tests, and also took the tests in order to be a reliability-tested possible fill-in for the intervention. The project leader has ten years of clinical experience assessing infant cranial asymmetry and instructs child health care nurses about cranial asymmetry detection. The project leader therefore served as reference rater in the infant test.

DESIGN

Training of Assessors

The project leader first organized an education where assessors were specifically trained to assess cranial asymmetry of infants using Cranial Technologies' Severity Assessments. The education consisted of two lessons each lasting two and one half hours. Each lesson included a power point presentation and verbal instructions. The Severity

Assessments were explained in detail. The assessors were given the opportunity to ask questions, to discuss within the group, and to practice.

Reliability Test

A reliability test was devised in order to evaluate intra-rater and inter-rater agreement after assessors had been trained. This consisted of a photograph test and an infant test. The photograph test was designed after Öhman's photograph test in a reliability study of physical therapists assessing plagiocephaly [21]. Fifty colored photographs of infant heads were selected for the photograph test, coded, and sent to assessors. Cranial asymmetry was assessed according to the three sets of pictures on photograph test score sheets. The stepwise pattern of increasing asymmetry in each set of pictures guided the assessors in making decisions. They decided what type and what degree of asymmetry they perceived. Type of asymmetry in vertex view photographs was classified according to what assessors perceived as the most dominant type of asymmetry - *skewed* which refers to plagiocephaly or *flat* which refers to brachycephaly. While vertex view photographs allowed for different skew and flat interpretations, if there was asymmetry in lateral view photographs, it was always *flat* since a skew could not be detected in the lateral view photographs (Fig. **1**)**.**

Vertex view and lateral view photographs were rated for degree of asymmetry using scores from 0-3. After completing the test the first time, the photographs and score sheets were returned to the project leader. The assessors were retested using the same procedure. To avoid recall bias, there was at least one week time lapse between the two occasions and the 50 photographs were re-coded and in a different order on the second occasion.

Since assessing cranial asymmetry in photographs and in infants is not an identical procedure, an infant test was included in order to examine how well assessors performed in a clinical setting. The infant test involved assessing cranial asymmetry in six infants according to the six sets of pictures from the Severity Assessments that were copied onto the infant test score sheets. An assessor was therefore expected to assess six features and record six scores for each infant. Assessors went into the room one at a time, assessed an infant, and then left the room before the next assessor went into the room. Type of cranial asymmetry could be judged as skewed, flat, a combination of the two, or no asymmetry. Degree of asymmetry was rated using scores from 0-3. Furthermore, half steps, scores half way between 1 and 2 for example, were permitted to increase precision of the judgments, since a lot more information is available when assessing and palpating the heads of infants compared to assessing 2-dimensional photographs.

STATISTICAL METHODS AND DATA ANALYSIS

The Photograph Test

Descriptive statistics included frequencies, means, percentages, and agreement matrices. Intra-rater and inter-rater agreement based on degree of asymmetry were analyzed by calculating the AC2 using quadratic weights. Observed values were then adjusted for chance agreement by subtracting the so called critical values in order to arrive at the perceived genuine agreement [22]. Landis and Koch's strength of agreement intervals for categorical data were chosen for interpretation of results [23].

The Infant Test

Descriptive statistics included frequencies and means. Inter-rater agreement based on degree of asymmetry was analyzed with the AC2 using quadratic weights, adjusted and interpreted as above.

Agreement Matrices

Agreement matrices were devised to illustrate assessors' agreement regarding both *type* and *degree* of asymmetry in photographs. The three types of asymmetry are labeled vertex view skew, vertex view flat, and lateral view flat, and they are arranged in that order in the columns and rows of a matrix. Four degrees of asymmetry, that is 0–3, are also included in the columns and rows, creating "four by four blocks" along the diagonal of a matrix.

Each intra-rater matrix illustrates one assessor's test-retest agreement when the same 50 photographs were

Fig. (1). Photographs of infant cranial shape similar to those used in the photograph test (parental consent obtained).

assessed on two occasions. This resulted in 50 observations in each intra-rater matrix. Each inter-rater matrix illustrates the summed inter-rater agreement of the 15 inter-rater pairs when six assessors assessed the 50 photographs on one of the occasions. The 15 inter-rater pairs are created from all the different assessor combinations when six assessors are tested on one occasion and intra-rater combinations are removed. The 15 inter-rater pairs multiplied by the 50 assessed photographs resulted in 750 inter-rater observations each time the test was taken.

Perfect agreement is expressed as frequencies on the diagonal of a matrix. Intra-rater agreement was perfect when an individual recorded the exact same judgment regarding both *type and degree* of asymmetry for a particular photograph both times that individual took the test. Inter-rater agreement was perfect when an inter-rater pair recorded the exact same judgment regarding both type and degree of asymmetry for a particular photograph in the first series of photographs – labeled Test in the matrices, or the second series of photographs – labeled Retest in the matrices.

Disagreements regarding *degree* of asymmetry are denoted as frequencies in non-diagonal cells where increased distance from the diagonal of a matrix indicates stronger disagreement.

"*Misclassifications*", disagreements regarding *type* of asymmetry, are denoted as frequencies in the area outside the four by four blocks. An example of an intra-rater "misclassification" is if an individual assessed a photograph as skewed in the Test and flat in the Retest. An example of an inter-rater "misclassification" is if Assessor A assessed photograph Z as skewed while Assessor B assessed photograph Z as flat in the Retest. On the other hand, if a photograph was assessed as no asymmetry once and mild asymmetry once, the observation is considered a disagreement in *degree* of asymmetry, not a "misclassification".

Agreement Measure

An agreement coefficient named "Agreement Coefficient 2" and denoted AC2 was used to calculate agreement based on *degree* of asymmetry. This agreement measure allows for multi-raters and ordered categorical data, weighs disagreements, and adjusts for chance agreement to arrive at the perceived genuine agreement [22]. Furthermore, it limits chance agreement probability to a maximum of 50 % [24].

Each estimated agreement coefficient was adjusted for possible sampling errors before interpretation. Adjustment was done by subtracting a so called critical value from the observed value to avoid agreement values inflated by chance. The adjustments made depended on the number of subjects, raters, and response categories used in a situation when everything else is equivalent. Adjusted coefficients were then interpreted according to pre-defined standards that are widely used in agreement statistics but arbitrary [22]. Landis and Koch's strength of agreement intervals for categorical data were selected for our interpretations [23].

ETHICAL CONSIDERATIONS

Approval from the Regional Ethics Review Board in Gothenburg, Sweden was obtained for the project (Dnr: 418-11). Written informed parental consent was obtained for each infant. If any significant asymmetry was detected in infants, parents were offered professional advice.

Photographs were selected among publicly published photographs on various plagiocephaly and brachycephaly websites. No child could be identified in the photographs by these assessors. Furthermore, photographs were only used for the photograph test, and assessors returned all photographs to the project leader after each test was taken. Photographs were not distributed or used in any other way.

RESULTS

The Photograph Test

Intra-rater agreement is described in six intra-rater matrices (Table **1**). Assessors are denoted A through F in Table **1** and the percentages of perfect agreement were 72, 64, 74, 84, 78 and 64 respectively, mean 73. In other words, nearly three quarters of all the intra-rater observations had perfect agreement regarding *type and degree* of asymmetry. Disagreements regarding *degree* of asymmetry are infrequent and adjacent to the diagonals of matrices in all but three instances, indicating most agreement was minor. The sporadic "misclassifications" are seen in areas outside the four by four blocks. Four vertex view photographs accounted for all of the "misclassifications".

Inter-rater agreement is described in two inter-rater matrices (Table **2**). Inter-rater agreement in the Test and the Retest appear to be quite similar. Of the 750 inter-rater observations in the Test, 491 are on the diagonal (65% of perfect agreement). There are 251 degree of asymmetry disagreements (33% of responses) and 217 of these (86%) are adjacent to the diagonal, indicating most disagreement was minor. There are 25 observations outside the four by four blocks, so these are considered "misclassifications" (3% of responses), and the same four vertex view photographs as before accounted for all of them. Of the 750 inter-rater observations in the Retest, 473 are on the diagonal (63% of perfect agreement). There are 269 degree of asymmetry disagreements (36% of responses) and 235 of these (87%) are adjacent to the diagonal, indicating most disagreement was minor. There are 14 observations outside the four by four blocks, so these are considered "misclassifications" (2% of responses), and the same four vertex view photographs as before accounted for all of them.

Strength of agreement based on degree of asymmetry was substantial for the intra-raters when analyzed with the AC2 using quadratic weights, adjusted (Table **3**), and then interpreted according to Landis and Koch's intervals (Table **4**). There was substantial strength of agreement for each of the six assessors. Mean adjusted intra-rater AC2 corrected for chance agreement was 0.69 [0.63; 0.76].

Strength of agreement based on degree of asymmetry was also substantial for the inter-raters when analyzed with the AC2 using quadratic weights, adjusted, and interpreted according to Landis and Koch's intervals. The inter-rater AC2 was 0.82 in the Test and 0.81 in the Retest. The adjusted inter-rater AC2s corrected for chance agreement were 0.72 and 0.71 respectively (Table **3**). There were no missing values in the photograph test.

Table 1. Frequencies of intra-rater agreement of cranial asymmetry in 50 photographs and sorted by type of asymmetry as judged by 6 assessors tested and retested one week later.

Assessor A		**RETEST**											
		Vertex view skew				Vertex view flat				Lateral view flat			
TEST		0	1	2	3	0	1	2	3	0	1	2	3
Vertex view skew	0	3											
	1												
	2		1	1	2								
	3			1	6			1					
Vertex view flat	0					3							
	1					1							
	2			1			1						
	3								1				
Lateral view flat	0									11	1		
	1									1	6		
	2										2	5	2
	3												3

Assessor B		**RETEST**											
		Vertex view skew				Vertex view flat				Lateral view flat			
TEST		0	1	2	3	0	1	2	3	0	1	2	3
Vertex view skew	0	3	1										
	1		1					1					
	2		2	2	2								
	3		1	1	3								
Vertex view flat	0					3							
	1												
	2							1					
	3								1				
Lateral view flat	0									12	2		
	1									3	5	2	
	2										1	3	2
	3												1

Assessor C		**RETEST**											
		Vertex view skew				Vertex view flat				Lateral view flat			
TEST		0	1	2	3	0	1	2	3	0	1	2	3
Vertex view skew	0	4											
	1	1	2										
	2		3	3				1					
	3			1	3								
Vertex view flat	0					4							
	1												
	2												
	3								1				
Lateral view flat	0									9	1		
	1									2	8	2	
	2										1	3	
	3											1	4

Assessor D		**RETEST**											
		Vertex view skew				Vertex view flat				Lateral view flat			
TEST		0	1	2	3	0	1	2	3	0	1	2	3
Vertex view skew	0	4											
	1	1	1				1						
	2		1	3	1								
	3				3								
Vertex view flat	0					4							
	1		1										
	2							2					
	3								1				
Lateral view flat	0									17			
	1										5		
	2											5	1
	3											2	1

Assessor E		**RETEST**											
		Vertex view skew				Vertex view flat				Lateral view flat			
TEST		0	1	2	3	0	1	2	3	0	1	2	3
Vertex view skew	0	2											
	1	1											
	2	2		4									
	3				7								
Vertex view flat	0					2							
	1												
	2						2						
	3							1					
Lateral view flat	0									15	1		
	1									1	4	1	
	2											3	2
	3												4

Assessor F		**RETEST**											
		Vertex view skew				Vertex view flat				Lateral view flat			
TEST		0	1	2	3	0	1	2	3	0	1	2	3
Vertex view skew	0	1	5										
	1		1										
	2			3	2								
	3			1	3								
Vertex view flat	0					1	1						
	1						1		1				
	2												
	3												
Lateral view flat	0									17			
	1									2	2	2	
	2											1	4
	3												3

Degree of asymmetry: 0 = none, 1 = mild, 2 = moderate and 3 = severe.

Perfect agreement is expressed as frequencies on the diagonal. Disagreements regarding degree of asymmetry are expressed as frequencies in non-diagonal cells where increased distance from the diagonal indicates stronger disagreement. Frequencies in areas outside four by four blocks indicate a photograph was judged as skewed one time and flat the other time.

Table 2. Frequencies of inter-rater agreement sorted by type of asymmetry in 50 photographs when 6 assessors judged cranial asymmetry on 2 occasions.

TEST		ASSESSOR 2											
		Vertex view skew				Vertex view flat				Lateral view flat			
ASSESSOR 1		0	1	2	3	0	1	2	3	0	1	2	3
Vertex	0	46*	4	4									
view	1	4	8	3	2		2	4					
skew	2	8	3	51	14		3	2					
	3		6	21	54		1	3					
Vertex	0					46*							
view	1			1		5							
flat	2		5	4		4	3	5					
	3						5		10				
Lateral	0									177	19		
view	1									53	56	18	
flat	2									4	28	50	16
	3											10	34

RETEST		ASSESSOR 2											
		Vertex view skew				Vertex view flat				Lateral view flat			
ASSESSOR 1		0	1	2	3	0	1	2	3	0	1	2	3
Vertex	0	42*	16										
view	1	7	13	12	6		2	1					
skew	2		10	28	9			4					
	3	2	10	14	62								
Vertex	0		1			42*	2						
view	1	1					7	1					
flat	2		3	2		2	9	4	1				
	3							4	10				
Lateral	0									180	23		
view	1									58	41	19	4
flat	2									9	8	39	25
	3										1	11	47

Degree of asymmetry: 0 = none, 1 = mild, 2 = moderate and 3 = severe.

Perfect agreement is expressed as frequencies on the diagonal. Disagreements regarding degree of asymmetry are expressed as frequencies in non-diagonal cells where increased distance from diagonal indicates stronger disagreement. Frequencies in areas outside four by four blocks indicate a photograph was judged as skewed by one assessor and flat by another.

*Counts included in 2 cells when a photograph was judged to have no asymmetry since that indicates both 0 skew *and* 0 flat agreement. These counts are only used once in calculations.

The Infant Test

The six four-month-old infants that were assessed on the same day in the same clinic are described in Table **5** according to background information suspected to be determinants of NSP. The measures we considered are typically found in healthy infants. They were present in participants just as they are present in the general infant population, and therefore, these infants were deemed suitable for the test.

Table **6** describes inter-rater agreement based on *type* of cranial asymmetry. When asymmetry was present according to the reference rater, the assessors *detected* it, but there was no complete agreement among assessors for any of the five infants with asymmetry. In Table **7**, assessors' *detection* of asymmetry corresponded with the reference rater in 23 of 24 instances, which is nearly perfect. The only aberration was that Rater 4 missed asymmetry in one infant. Inter-rater agreement based on *degree* of asymmetry was analyzed with AC2 using quadratic weights and found to be 0.83 [0.70; 0.97]. After correction for chance agreement, the adjusted AC2 was 0.73 [0.60; 0.87]. This indicates substantial strength of agreement based on degree of asymmetry when interpreted according to Landis and Koch's intervals.

DISCUSSION

The study shows that the assessors were reliable and inter-changeable when using the Severity Assessments as a tool to estimate infant cranial asymmetry. Furthermore, their ability to *detect* asymmetry in the clinical setting was excellent. We strove for a thorough analysis since evaluating reliability of assessors in this study was not straightforward for several reasons.

The scales in Cranial Technologies' Severity Assessments are linear, yet we believe the difference between no asymmetry and mild asymmetry represents a small difference, while the difference between mild and moderate is an important difference because moderate asymmetry includes secondary asymmetries [16, 17]. The asymmetry progression between moderate and severe is probably much more important, because, as we see it, increasing skull alterations involve increasing consequences for the infant. Furthermore, asymmetry progression is actually on a continuum, not in steps. Because of our non-linearity assumption concerning increasing asymmetry, we chose to use quadratic weights when analyzing agreement based on degree of asymmetry. We assumed that a quadratic increase best reflects the increasing asymmetry since we were unable to ascertain this. This information is not available as far as we know.

Translating agreement into a valid and fair single measure is a complex undertaking. The appropriate agreement measure needs to be chosen. Cohen's kappa only works for two raters while Fleiss' kappa works for multi-raters but does not consider weighing disagreement [25]. AC2 allows for multi-raters and ordered categorical data, weighs disagreements, and adjusts for chance agreement. Furthermore, the AC2 limits chance agreement probability to a maximum of 50%, a reasonable value, while kappa statistics do not [24]. It was therefore deemed the only appropriate agreement measure for this study. The AC2 produces more robust measurements due to the way the

Table 3. Agreement coefficients using AC2 with quadratic weights and corrected for chance agreement when 6 assessors judged degree of cranial asymmetry in 50 photographs on two occasions.

Assessors	AC2 with Quadratic Weights [95% CI]	AC2 [95% CI] Corrected for Chance Agreement
Assessor A	0.90 [0.85; 0.95]	0.70 [0.65; 0,75]
Assessor B	0.84 [0.76; 0.92]	0.64 [0.56; 0.72]
Assessor C	0.91 [0.86; 0.96]	0.71 [0.66; 0.76]
Assessor D	0.96 [0.92; 0.99]	0.76 [0.72; 0.79]
Assessor E	0.88 [0.79; 0.97]	0.68 [0.59; 0.77]
Assessor F	0.86 [0.78; 0.94]	0.66 [0.58; 0.74]
Intra-rater mean	0.89 [0.83; 0.96]	0.69 [0.63; 0.76]
Inter-raters test 1	0.82 [0.74; 0.91]	0.72 [0.64; 0.81]
Inter-raters test 2	0.81 [0.73; 0.89]	0.71 [0.63; 0.79]

The heading Assessors in column 1 refers to 6 individuals and mixed tables – the intra-rater mean and the inter-rater pairs created the first time and second time the test was taken. Column 2 includes observed values. Column 3 includes values corrected for chance agreement to arrive at perceived genuine agreement.

agreement values are calculated. The next step is to interpret these values. We chose from a number of different value systems, all quite arbitrary by definition [22], and we decided on Landis and Koch since it is widely used, well known and generally accepted. On the other hand, if our degree of asymmetry results had not been adjusted, they would all have been interpreted as "nearly perfect" according to Landis & Koch's strength of agreement intervals.

Table 4. An adaptation from Landis & Koch's intervals when interpreting strength of agreement using kappa statistics.

Agreement Statistic	Strength of Agreement
< 0.0	Poor
0.00 to 0.20	Slight
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Substantial
0.81 to 1.00	Almost Perfect

Only one trait should be analyzed at a time in agreement statistics. We examined two traits – *degree* of asymmetry that could be analyzed with an agreement measure and *type* of asymmetry that could not be analyzed. We therefore devised an agreement matrix which enables a graphic description of agreement based on two traits. This is an idea that we find provides a richer description of the agreement.

There were several aspects to consider in the descriptions of agreement. Information available in photographs is limited, and vertex view photographs leave room for different interpretations. Some photographs were not as straightforward as others. They were purposely chosen to encourage assessors to think more carefully. Sometimes it was difficult for assessors to discern if vertex view photographs depicted skewed or flat cranial asymmetry. Due to the partial information, there was some educated guessing, and consequently, we would expect some variability in the photograph test. This potentially explains the "misclassifications" seen outside the four by four blocks in the matrices. As it turned out, the same four vertex view photographs accounted for all of these deviations.

Table 5. Frequencies and means of background information on 6 infants judged for cranial asymmetry on the same day at one clinic.

	Frequency	Mean
Infant Characteristics		
Males	2	
Females	4	
Twins	2	
Mean age at assessment		15.3 weeks
Birth Related Factors		
Firstborn	3	
Normal delivery	5	
Vacuum extraction	1	
Mean birth weight		3,284 grams
Mean gestational age		38.8 weeks
Other Suspected Determinants		
Cranial flat spot at birth	0	
Exhibited side preference	3	
Only bottle fed	3	
Always placed supine	2	
No infant pillow when sleeping supine	2	

Information includes characteristics and factors potentially related to cranial asymmetry.

Since a photograph test is probably not the best way to judge reliability of assessors, an infant test was included. Assessing infants allows for palpation and also assessment of more asymmetry features. However, assessing infants complicates decision making. Infants cannot be depended on to keep still, and it is a challenge to assess infants that do not cooperate. Infants' behavior and/or position may have differed during the 5 consecutive assessments each infant

Table 6. Classification of cranial asymmetry seen in rows for each of 6 infants judged on the same day at the same clinic by 4 assessors and a reference rater.

Infant	Assessor				Reference Rater
	A1	A2	A3	A4	
1	b	b	b	b	p+b
2	p+b	p+b	p	p	p+b
3	p+b	p	p+b	p	p+b
4	p+b	b	p+b	b	b
5	p+b	b	p+b	0	p+b
6	0	0	0	0	0

Key: p = plagiocephaly, b = brachycephaly, p+b = both, and 0 = no flattening.

Table 7. Frequencies of four assessors' agreement with the reference rater regarding presence of cranial asymmetry in 6 infants assessed one day at the same clinic.

Assessor	Asymmetry	Asymmetry by Reference Rater	
		No	Yes
A	**No**	1	0
	Yes	0	5
B	**No**	1	0
	Yes	0	5
C	**No**	1	0
	Yes	0	5
D	**No**	1	1
	Yes	0	4

was subjected to in the infant test, which may have affected decisions. Sometimes decision making gets rushed. Furthermore, the human eye cannot readily discern millimeter differences in the cut-off points in the sets of pictures in the Severity Assessments, and an individual's features do not always correspond with the template. In addition, different infant characteristics leave room for different interpretations. Therefore, consistency becomes complicated. The complexity of assessing live infants is illustrated in the following observations an assessor spontaneously wrote: *The whole procedure with assessments is certainly complex. There is so much that happens during the measurement situation, parents, siblings, children, and the relationship one gets. Lively, alert, cooperative, suspicious, tired, children with long hair, etcetera, children who look at us in different ways. Staff, the clinics, or home visits.* Nevertheless, it seems as though the assessors performed just as well in a clinical setting, because the inter-rater AC2s in this study are nearly identical in the photograph test and the infant test.

An agreement matrix was not useful in the infant test due to the numerous rating options and the small sample. Instead, we examined the clinical agreement of the assessors, similar to the concepts of sensitivity and specificity (Table **6**). When comparing these 24 asymmetry/no asymmetry decisions to the reference rater, asymmetry was missed only once by one assessor and there was no over-diagnosing. With the simplified agreement criterion "asymmetry" or "no asymmetry", clinical performance appeared to be excellent.

Physical findings often rely on some degree of subjective interpretation by observers [13]. Some possible systematic variation was noticed in the data. This could be due to different attitudes and/or different ways of reasoning. These are subjective influences that cannot be removed by better planning. Nevertheless, as Bland and Altman wrote, " We should ask if agreement is good enough for a particular purpose, not whether it conforms to some absolute, arbitrary criterion" [26]. This pragmatic attitude is encouraging, because while precision instruments do not seem appropriate for cranial asymmetry screening during child health visits, the Severity Assessments seem to be both appropriate and useful.

Also encouraging for our project is that specially trained pediatric nurse practitioners working in a cranial facial program in the United States were found to be effective and safe when assessing NSP and screening for craniosynostosis [27]. That study is not directly comparable to what we did in our project because these were nurse practitioners who worked in a craniofacial center and were trained by cranial facial surgeons. These nurse practitioners screened for craniosynostosis and followed up NSP cases, whereas the focus of our project is to develop effective prevention strategies by providing an NSP education for the child health nurse clinicians. The child health nurse clinicians in Sweden are required to be either public health nurse specialists or pediatric nurse specialists. Nevertheless, these are both examples of the increasing involvement and expanding responsibility of nurses working with NSP. This reliability study is one step in our NSP prevention methodology development.

STRENGTHS AND LIMITATIONS OF THE STUDY

The study has several strengths. The design includes both a photograph test and a live subject test. There is no missing data in the photograph test. An agreement matrix was devised to provide a richer description of the agreement. The selected agreement measure was appropriate for the situation and the choice of weights was carefully deliberated. The AC2s were adjusted for chance agreement to avoid results that are artificially inflated by sampling errors. When interpreted, results indicate that strength of agreement based on degree of asymmetry was substantial, one interval less than "almost perfect". We have reported subjective influences conscientiously.

A limitation is that a photograph test is not the optimal way to judge reliability of assessors whose intended task is to assess cranial asymmetry in live infants. Another limitation is that the sample in the infant test was small.

CONCLUSION

Assessors' agreement when assessing infant cranial asymmetry was substantial. They can therefore be considered reliable interchangeable assessors in the intervention study.

CLINICAL IMPLICATIONS

Assessors' role in this study – being specifically trained and then assessing infant cranial asymmetry – represents the expanded role of our trained child health nurse clinicians. If we consider the assessors as proxies for these nurses, their performance in this study can be extended to the larger clinical context. The original idea, that specifically training child health nurse clinicians to assess infant cranial asymmetry might be helpful for early detection, seems to work. Likewise, it could be helpful to specifically train other professional clinicians working with infants how to assess cranial asymmetry.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

The authors would like to thank Anna Öhman, PhD, for helping design the photograph test, for manuscript suggestions and for dedication to plagiocephaly prevention. The authors would like to thank Professor Emerita Ingela Krantz for manuscript suggestions. The study was supported by the Sahlgrenska Academy at the University of Gothenburg, the Skaraborg Institute in Skövde, Sweden, Primary Care Skaraborg, and the Queen Silvia Children's Hospital Research Fund.

REFERENCES

[1] Littlefield TR, Saba NM, Kelly KM. On the current incidence of deformational plagiocephaly: An estimation based on prospective registration at a single center. Semin Pediatr Neurol 2004; 11: 301-4.

[2] Wilbrand JF, Schmidtberg K, Bierther U, *et al*. Clinical classification of infant nonsynostotic cranial deformity. J Pediatr 2012; 161: 1120-5.

[3] Argenta L, David L, Wilson J, Bell W. An increase in infant cranial deformity with supine sleeping position. J Craniofac Surg 1996; 7: 5-11.

[4] Biggs WS. The 'epidemic' of deformational plagiocephaly and the American Academy of Pediatrics' response. JPO 2004; 16: S5-S8.

[5] Kane AA, Mitchell LE, Craven KP, Marsh JL. Observations on a recent increase in plagiocephaly without synostosis. Pediatrics 1996; 97: 877-85.

[6] Mawji A, Vollman AR, Hatfield J, McNeil DA, Sauve R. The incidence of positional plagiocephaly: A cohort study. Pediatrics 2013; 132: 298-304.

[7] Lauritzen C, Tarnow P. Assymetric skull? Early correct diagnosis is a must! Positional skull deformities can be avoided. Lakartidningen 1999; 96: 1447-53.

[8] van Wijk RM, van Vlimmeren LA, Groothuis-Oudshoorn CG, Van der Ploeg CP, Ijzerman MJ, Boere-Boonekamp MM. Helmet therapy in infants with positional skull deformation: randomised controlled trial. BMJ 2014; 348: g2741.

[9] Hutchison BL, Stewart AW, Mitchell EA. Deformational plagiocephaly: A follow-up of head shape, parental concern and neurodevelopment at ages 3 and 4 years. Arch Dis Child 2011; 96: 85-90.

[10] Roby BB, Finkelstein M, Tibesar RJ, Sidman JD. Prevalence of positional plagiocephaly in teens born after the "back to sleep" campaign. Otolaryngol Head Neck Surg 2012; 146: 823-8.

[11] Lennartsson F. Developing guidelines for child health care nurses to prevent nonsynostotic plagiocephaly: Searching for the evidence. J Pediatr Nurs 2011; 26: 348-58.

[12] Lennartsson F. Testing guidelines for child health care nurses to prevent nonsynostotic plagiocephaly: A Swedish pilot study. J Pediatr Nurs 2011; 26: 541-51.

[13] Viera AJ, Garrett JM. Understanding interobserver agreement: The kappa statistic. Family Med 2005; 37: 360-3.

[14] Glasgow TS, Siddiqi F, Hoff C, Young PC. Deformational plagiocephaly: Development of an objective measure and determination of its prevalence in primary care. J Craniofac Surg 2007; 18: 85-92.

[15] Robinson S, Proctor M. Diagnosis and management of deformational plagiocephaly. JNS: Pediatrics 2009; 3: 284-95.

[16] Argenta L, David L, Thompson J. Clinical classification of positional plagiocephaly. J Craniofac Surg 2004; 15: 368-72.

[17] Hylton-Plank L. The presentation of deformational plagiocephaly. JPO 2004; 16: 28-30.

[18] CranialTechnologies. Online head shape assessment [cited 2014 December 11]; Available from: http://www.cranialtech.com/index.php?option=com_content&view=article&id=108&Itemid=78.

[19] Feijen M, Schuckman M, Habets E, van der Hulst R. Positional plagiocephaly and brachycephaly: Is there a correlation between subjective and objective assessment of cranial shape? J Craniofac Surg 2012; 23: 998-1001.

[20] Cavalier A, Picot MC, Artiaga C, *et al*. Prevention of deformational plagiocephaly in neonates. Early Hum Dev 2011; 87: 537-43.

[21] Ohman A. The inter-rater and intra-rater reliability of a modified "severity scale for assessment of plagiocephaly" among physical therapists. Physiother Theory Pract 2012; 28: 402-6.

[22] Gwet K. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters. 3rd ed. Gaithersburg: Advanced Analytics, LLC 2012; pp. 121-47.

[23] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-74.

[24] Gwet K. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Series: Stat Methods Inter-Rater Reliabil Assess 2002; 1: 1-5.

[25] Uebersax J. Statistical methods for rater and diagnostic agreement [cited 2014 October 2]; Available from: http://www.john-uebersax.com/stat/agree.htm

[26] Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med 1990; 20: 337-40.

[27] Kuang AA, Bergquist C, Crupi L, Oliverio M, Selden NR. Effectiveness and safety of independent pediatric nurse practitioners in evaluating plagiocephaly. Plast Reconstr Surg 2013; 132: 414-8.

Received: March 20, 2015 Revised: June 2, 2015 Accepted: June 15, 2015