Evaluation of Interobserver and Intraobserver Differences of Graf Method in Developmental Hip Dysplasia

Gelişimsel Kalça Displazisinde Graf Yönteminin Gözlemciler Arası ve Gözlemci İçi Farklılıklarının Değerlendirilmesi

Halil Dogruel[1], Alim Can Baymurat[2], Ismail Cengiz Tuncay[3], Hakan Atalar[2]

[1] Etimesgut Sehit Sait Erturk State Hospital, Department of Orthopaedics and Traumatology, Ankara, Turkiye
[2] Gazi University Medical Faculty, Department of Orthopaedics and Traumatology, Ankara, Turkiye
[3] Başkent University Ankara Hospital, Department of Orthopaedics and Traumatology, Ankara , Turkiye

ABSTRACT

Purpose: The aim of this study was to evaluate possible differences in hip ultrasonography (US) results between physicians working in different medical centers.
Method: In this study, a total of 117 horizontal US images representing all sonographic types of developmental dysplasia of the hip (DDH) were utilized. Four experienced researchers independently measured the 117 US images at different times. The results obtained by each researcher were documented separately, including alpha and beta angles, Graf types. All of the collected data were analyzed statistically to assess for interobserver and intraobserver variability.
Results: The study found that the average change between the alpha angles ranged from a minimum of 1 to a maximum of 4. The mean change between the beta angles was wider, ranging from a minimum of 1.8 to a maximum of 8.2. To evaluate the differences between the Graf hip typologies, paired groups were formed and Cohen's Kappa method was used. For the first group k= 0.661, for the second group k= 0.671, for the third group K= 0.647, for the fourth group k= 0, 718, k= 0.717 for the fifth group and k= 0.637 for the sixth group. Interobserver Kappa evaluation results (k=0.647) showed moderate and significant agreement.
Conclusion: The results revealed a moderate to substantial level of agreement between the researchers. Based on these findings, it was concluded that the use of US for screening and follow-up of the Graf hip typing method should be performed by experienced professionals.

Keywords: Developmental hip dysplasia, ultrasonography, Graf method, interobserver, intraobserver

ÖZET

Amaç: Bu çalışmanın amacı, farklı sağlık kuruluşlarında çalışan hekimler arasında kalça ultrasonografisi sonuçlarındaki olası farklılıkları değerlendirmektir.
Yöntem: Çalışmada, gelişimsel kalça displazisinin tüm sonografik tiplerini temsil eden toplam 117 yatay US görüntüsü kullanıldı. Dört deneyimli araştırmacı birbirlerinden bağımsız olarak117 US görüntüsünü farklı zamanlarda değerlendirdi. Her bir araştırmacı tarafından elde edilen alfa, beta açıları ve Graf tipleme sonuçları ayrı ayrı belgelendi. Toplanan tüm verilerin, gözlemciler arası ve gözlemci içi değişkenliği değerlendirmek için istatistiksel analiz yapıldı.
Bulgular: Çalışma, alfa açıları arasındaki ortalama değişimin minimum 1 ile maksimum 4 arasında değiştiğini ortaya koydu. Beta açıları arasındaki ortalama değişim ise daha genişti ve minimum 1,8 ile maksimum 8,2 arasında olduğu görüldü. Graf kalça tipolojileri arasındaki farkları değerlendirmek için eşleştirilmiş gruplar oluşturularak Cohen's Kappa yöntemi kullanıldı. Birinci grup için k= 0.661, ikinci grup için k= 0.671, üçüncü grup için k= 0.647, dördüncü grup için k= 0, 718, beşinci grup için k= 0.717 ve altıncı grup için k= 0.637 olarak bulundu. Gözlemciler arası Kappa değerlendirme sonuçları (k=0.647) orta düzeyde ve anlamlı bir uyum gösterdi.
Sonuç: Çalışmanın sonucunda, araştırmacılar arasında orta ila yüksek düzeyde bir uyum görülmüştür. Bulgulardan yola çıkarak, Graf kalça tipleme yönteminin tarama ve takibi için US kullanımının deneyimli profesyoneller tarafından gerçekleştirilmesi gerektiğinin kanaatine varılmıştır.

Anahtar Sözcükler: Gelişimsel kalça displazisi, ultrasonografi, Graf yöntemi, gözlemciler arası, gözlemci içi

ORCIDs: H.D.0000-0002-6031-7064,A.C.B.0000-0002-0062-621X,I.C.T.0000-0001-5856-8895,H.A.0000-0002-0016-0552

doi:http://dx.doi.org/10.12996/gmj.2023.86

INTRODUCTION

Nowadays, hip ultrasonography has become widely used in screening, planning and follow-up of DHD. For the first time, Graf described the morphology of the newborn hip joint by ultrasonography, monitoring hip development, determination of hip dysplasia, ultrasonographic classification, and treatment planning (1).

In the literature, apart from the Graf (static, morphological) method, Terjesen (2), Harcke (3), Novick (4) (multiplanar dynamic) and Suzuki (5) defined their own ultrasonographic methods.

Difficulties in the application of Terjesen, Novick and Harcke's dynamic methods and the fact that Suzuki's method requires a long linear transducer have highlighted Graf's method of static evaluation. Today, Graf's US method is still the most widely used ultrasonographic diagnostic method with its easy applicability.

With this method, alpha and beta angles are measured by taking a central coronal cross-section through the acetabulum and femoral head (Figure 1 a,b), and acetabular development can be evaluated morphologically (6).

Due to the fact that it is a preventable disability, early diagnosis and treatment of developmental hip dysplasia is of utmost importance. In his articles, Ortolani emphasized the importance of early and very early diagnosis and tried to explain its effects on treatment (7). Recently, with hip ultrasonography, it has become possible to evaluate the hip joint, which has a cartilage structure, in the infantile period. It has been possible to initiate the treatment early with early diagnosis. However, the reliability, sensitivity and specificity of the methods guiding the diagnosis and treatment, as well as the accuracy of the diagnosis and the selected treatment are of utmost importance.

In our country, many studies are available on the sensitivity and specificity of the US method in DHD, the comparison of US with clinical examination, the frequency of DHD, and the effectiveness of ultrasonographic evaluation in treatment and follow-up (8-11). With the generalization of the Graf method, orthopedic surgeons, radiologists and pediatricians have also been included in the group of physicians who apply this method. Thus, it has become important to evaluate the results obtained by different physician groups applying the Graf technique, to evaluate the possible differences between the evaluations and the effects of the results on the treatment.

Possible differences in the results obtained may affect the reliability, sensitivity and specificity of the method. Therefore, an evaluation method was planned and measurements made on standardized US cross-sections, possible differences in measurements and their effects on treatment were investigated among physicians who frequently applied this method and worked in different centers.

MATERIAL and METHOD

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of our Institiuion (Date: 23.04.2022 / No: E-77082166-604.01.02-639365).

In the study, hip ultrasonography image outputs obtained with a standard horizontal cross-section and using standard Graff cushion were used. The all hip ultrasound images used in the study were obtained from the personal archive of the fourth author experienced in hip ultrasonography. Ultrasonographic outputs have been sampled to include all hip types defined by the Graf hip ultrasonographic method, and a total of 117 standardized horizontal US cross-sections were used for the study. In the measured ultrasonography, the name information about the patient was was hidden , but the age information was left visible so that type 2 a/b evaluation could be performed (Figure 1).
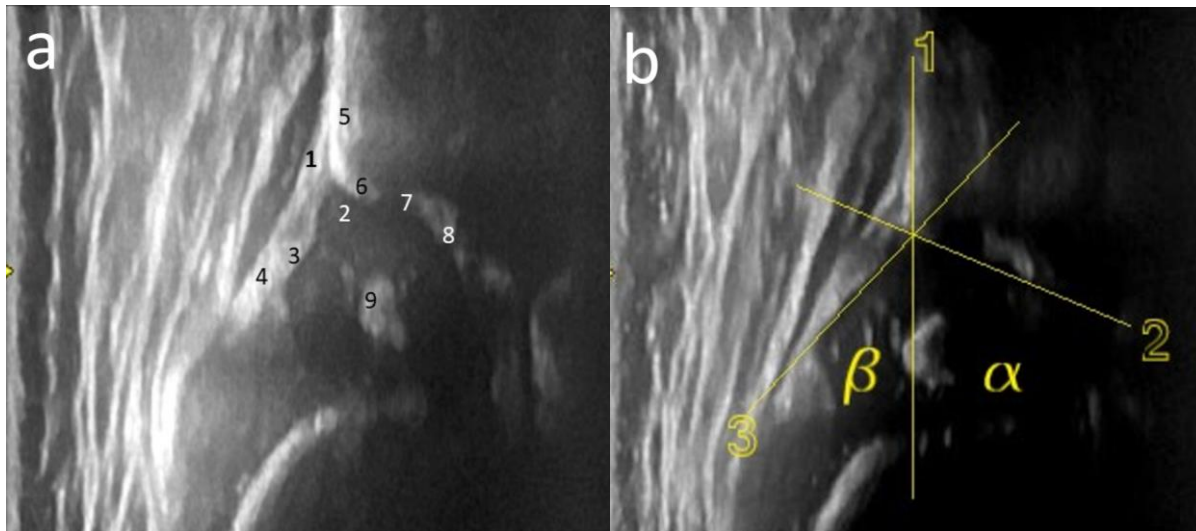

Figure 1. **a** Explanation of Graf ultrasonogram Image; **b** Alpha (α) and Beta (β) angles.
a 1- periosteum of the ilium and perichondrium, 2- cartilage acetabular roof, 3- acetabular labrum, 4- joint capsule, 5- ilium, 6- bone acetabular edge, 7- iliac bone, 8- deepest (inferior) point of the ilium , 9- femoral head .
b The alpha (α) angle refers to the angle between the line (1) drawn parallel to the ilium and the line (2) drawn from the deepest point of the ilium to the bone acetabular edge and allows us to evaluate the bone acetabular roof.
The beta (β) angle is the angle between the line (1) drawn parallel to the ilium and the line (3) drawn from the deepest point of the ilium to the acetabulary labrum and allows us to evaluate the cartilage acetabulary roof.

In addition to the evaluation of interobserver differences, with the help of these 117-image computer image processing programs, some changes were made on the image properties such as size, contrast, sharpness and brightness that would not affect the measurements, and self-repetitive images were obtained by trying to create the perception that the images were different. Only the author knows which of these images are identical to each other, and the researchers who participated in the study and measured were blinded. Thus, it was aimed to evaluate the intraobserver differences between the measurements. The images were given to the researchers in mixed order in special outputs in photo quality.

In order to prevent the negative effects of the technical differences in the measurements on the results of the study, all the researchers participating in the study measured using the same goniometer. The use of the same goniometer had a negative effect since it prolonged the duration of the study, but the use of a single goniometer was considered important in terms of preventing possible technical differences in the measurements. Each researcher was given enough time to make their measurements and evaluations easily.

The researchers who made the measurement were given a standard form and were asked to record the results they obtained from the measurements, alpha and beta angles, in this form. In addition, they were asked to perform a sonographic typing according to the Graf method by examining the ultrasonographic cross-sections morphologically and to record the follow-up and treatment recommendations accordingly.

The differences between the alpha and beta angles recorded statistically and the differences between the Graf types were tried to be examined as interobserver and intraobserver. As a statistical method, the Bland Altman statistical method, in which each measurement can be evaluated individually with the other, was used and possible differences and the reliability of the US measurement technique were tried to be evaluated.

However, since the standard deviations of the two results can be evaluated in this statistical analysis, the results of the measurements and evaluations made by the researchers in the study group were created in groups of two, and the measurements of each researcher were matched with each other and a large number of analyzes were made.

The data obtained from the measurements made by the researchers from the blinding and repeating US images were also evaluated using correlation methods for intraobserver evaluation.

Finally, Graf hip typing records, including the morphological evaluation requested from the researchers, treatment recommendations and possible differences between them were also evaluated using the Kappa statistical method. Statistical evaluation of all these methods was made using MedCalc 12.1.4.0 program.

RESULTS

A total of 117 standardized US cross-sections obtained by Graf method were evaluated and recorded by 4 researchers, alpha, beta angle measurements and US typing described by Graf method were recorded.

According to the results of the statistical examination, the correlation between the measurements was examined in order to determine the interobserver differences made by the researchers in the self-repeated images and Pearson's R values showing correlation were determined as 0.951 (sig p=0.000), 0.975 (sig p=0.000), 0.996 (sig p=0.000) and 0.997 (sig p=0.000), respectively. These results were revealed to be highly compatible with each other, and no intraobserver difference was observed between the measurements.

The statistical results of the examination of the interobserver differences and method safety between the researchers who made the measurements were based on the opinion that the measurements were compatible if the measured values had a normal distribution and the different measurement values were randomly distributed around zero and if 95% of them were between "d" -1.96s and "d" +1.96s " (12).

Bland Altman standard deviation distribution tables were created in binary groups and the results were given in table groups by comparing each researcher with another (Table groups 1 and 2).

Table Group 1 Results of examination of the measurements of alpha angles in Bland Altman binary groups
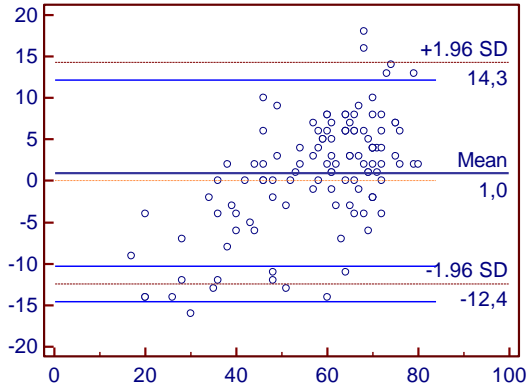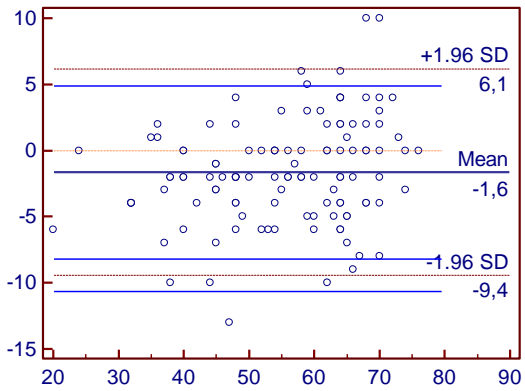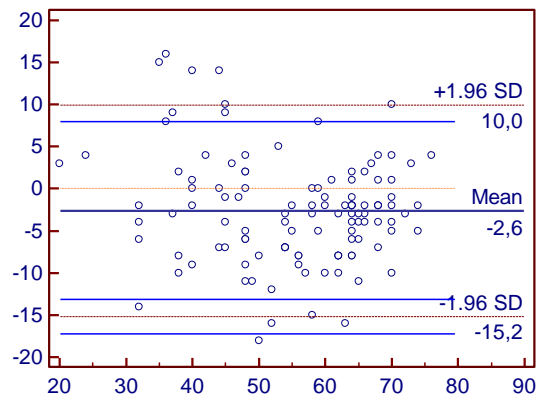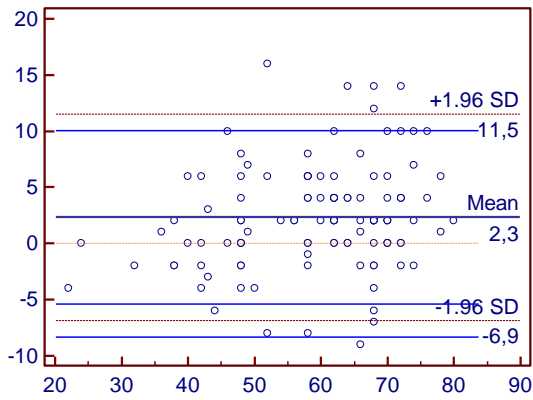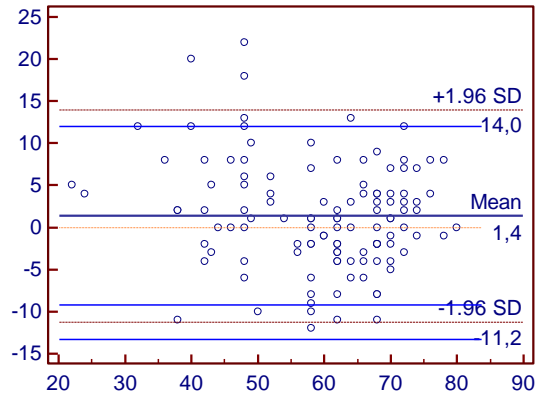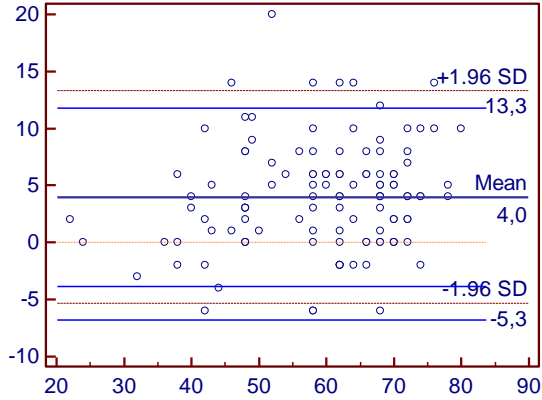
Table Group 2 Results of examination of the measurements of alpha angles in Bland Altman binary groups
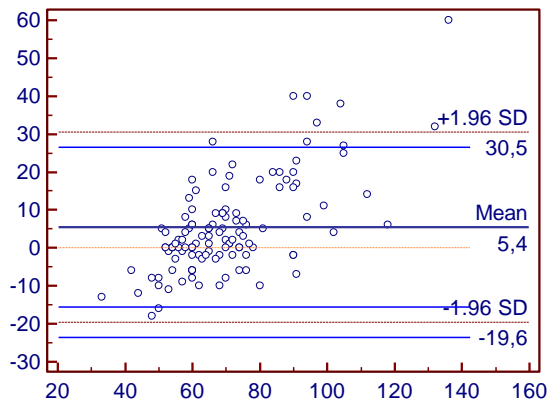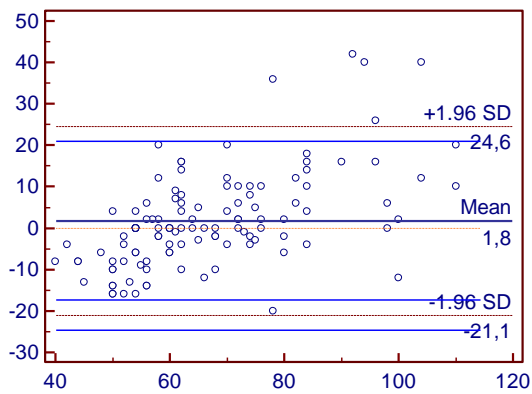
When the entire table groups are examined, it is observed that the differences between the measurements are randomly distributed around zero with a normal distribution, and 95% of them are between "d" -1.96s and d "+1.96s". Based on this observation, it can be said that the differences between the measurements remain within the acceptable standard deviation range. According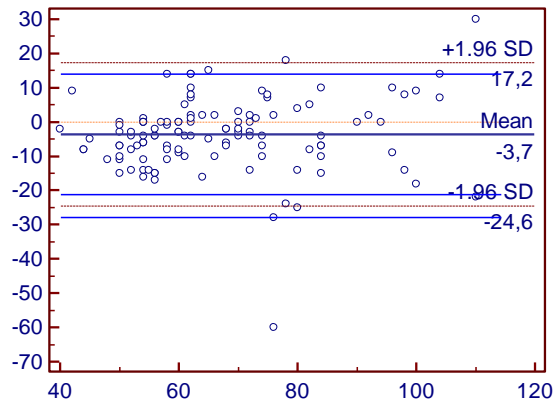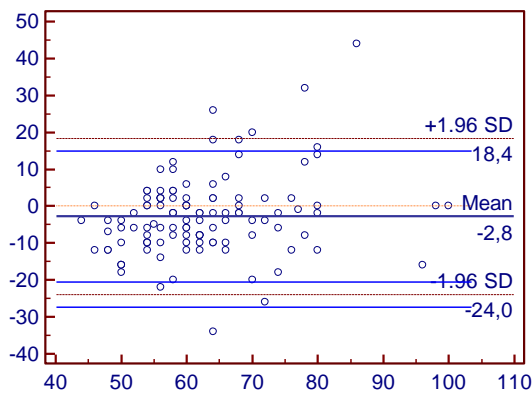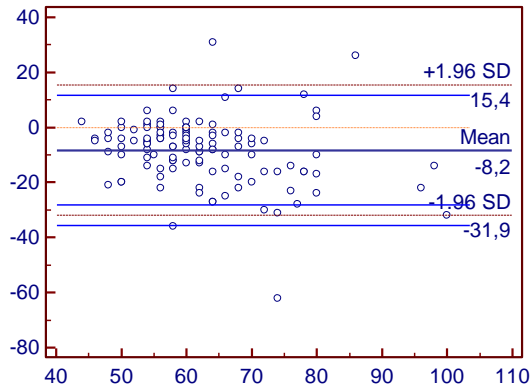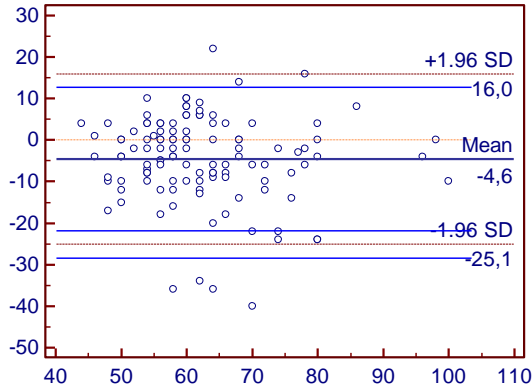ly, the mean change between alpha angles was at least 1 and at most 4 (range-15.2 - 14.3) (Table group 1). The mean change between beta angles was spread over a wider range and the mean change was measured in the range of at least 1.8 and at most 8.2 (range-31.9 – 30.5) (Table group 2).

When the differences between alpha angle measurements were examined, Cronbach's alpha value was calculated as 0.9736 and the 95% lower reliable limit was 0.9664. The change between the measurements occurred in the range of 0.0019-0.0133. The Cronbach's alpha value for beta angle measurements is 0.8992 and the 95% lower reliable limit is 0.8717. The change between measurements is in the range of 0.011-0.0536.

Since the differences between the typing were evaluated by conducting a morphological examination and there was no numerical value and it was a categorical evaluation, the correlation between the measurements and possible differences between the evaluations were evaluated by creating groups of 2 with Cohen's Kappa method.

All Kappa values were detected in the 95% safety range and the standard error occurred in the range of 0.036 to 0.040.

Kappa values of the binary groups were measured as, respectively, k= 0.661 (95% CI 0.582-0.739) for the first group, k= 0.671 (95% CI 0.600-0.742) for the second group, K= 0.647 (95% CI 0.564-0.730) for the third group, k= 0.718 (95% CI 0.651-0.786) for the fourth group, k= 0.717 (95% CI 0.647-0.787) for the fifth group, and k= 0.637 (95% CI 0.564-0.709) for the sixth group. Landis and Koch scale (13) was applied to evaluate Kappa results (Table 3). Interobserver Kappa evaluation results (k=0.647 – 0.787) according to this scale were moderate and substantial.

Table 3 Landis and Koch Kappa assessment scale

Kappa Statistics	Strength of Agreement ((Compatibility strength)
<0	Poor
0.01-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

DISCUSSION

In studies comparing hip ultrasonography with the follow-up and treatment of the physical examination of the hip joint, US methods have started to be widely used in reporting more reliable results than only physical examination and follow-up methods, screening of DHD with US methods, planning and follow-up of treatment. Although various US methods are also used, the US method (1) described by Graf has become more popular than others and has started to be used more than other US methods in screening and follow-up. The accuracy of diagnosis and treatment is extremely important for the treatment of developmental hip dysplasia, which is a preventable disability. For this reason, the reliability of this method, which guides diagnosis and treatment, should be at a high level. In our country, this method, described by Graf, started to be applied to all infants aged 6-8 weeks as a screening test applied to all newborns.

Alpha and beta angles from the hip joint are measured and treatment is planned by ultrasonographical typing according to the criteria described by the hip joint Graf with a morphological evaluation made in the US image. This method is widely used by orthopedic surgeons and radiologists in our country. Radiologists write and report Graf typing results by measuring alpha and beta angles according to the Graf method. Some orthopedic surgeons perform the US evaluation of the hip in infants themselves and plan treatment according to their results, while some surgeons decide to treat with radiology reports. Studies evaluating hip US efficacy in our country and in the world have also been published.

In the study conducted by Simona EA et al. (14) from Sweden, interobserver differences were evaluated between the groups consisting of orthopedics and radiology assistants who worked on the Graf method and principles, the participants were asked to blindly evaluate US outputs and measure alpha beta angles and US typing according to the method described by Graf. While recording the results, 4 result groups were created (type 1 mature, type 2 an immature, type 2b-2c-D dysplastic and type 3a-3b-4 dislocated) and they were asked to make an evaluation. In the study, Kappa values were determined in the range of 0.55-0.71 (moderate and substantial agreement). In the method we applied, the results were not grouped, but the agreement between experienced surgeons was evaluated in the study. The results obtained are at a similar level. The concordance between orthopedics and radiology assistants can be interpreted as the results being at a similar level to the experienced surgeons in our study, and the graph US method can be easily learned and practically used.

Barr-On E et al. (15) performed US evaluation with Graf method and obtained output on 75 infants with health proficiency and classified their results as normal (1a-2a) and pathological (2b-4) and the results obtained by two researchers were evaluated.

They made interobserver evaluation and interobserver evaluation by taking the same outputs more than once and measuring them again at different times. Interobserver agreement was observed as Kappa 0.50 and intraobserver agreement as kappa 0.61 in all results obtained. In the evaluation made by selecting only abnormal hip results, the consistency in the results has decreased considerably. Intraobserver Kappa was 0.41 and interobserver Kappa was 0.28 (poor agreement). The authors' comments were that although the same ultrasonography device and baby positioning device are used, low agreement shall require more standards in US imaging, which increases agreement in the future. In this study, both the performance of different researchers in US and the differences in measurements were combined, so the increasing negative effects decreased the agreement. We tried to take technical measures, including the use of the same goniometer, to prevent technical errors in our study and to make a more compatible evaluation. Increasing the technical standards in the US evaluation with the Graf technique shall positively affect the results.

In a study conducted by Ömeroğlu et al. (16) from our country, the Graf ultrasonography method was evaluated in detail and the evaluation results were examined in terms of interobserver and intraobserver differences. In the last year of the study, 22 researchers consisting of an orthopedic assistant, assistant professor and a professor were divided into 4 groups and the results of the morphological evaluations made for graph ultrasonography hip joint types and the agreement between these results were examined as interobserver and intraobserver with the measurements they made on the selected equal number of pathological and normal ultrasonography outputs standardized and repeated measurements at 4-week intervals. The mean intraobserver agreement rate between graph US types was 65% (40-90%) and kappa 0.52 (0.21-0.84), while the interobserver agreement rate was 51% (41-60%) and kappa 0.33 (0.21-0.43).

The fact that the agreement in Kappa values is less, the number of researchers making the measurement and the differences in experience can be thought to be, and the higher agreement in our study may be due to the fact that the measurements were made by 4 researchers with high experience, according to this study, the difference in our study is that the Alpha and Beta measurement values were evaluated by comparing each measurement with each other instead of an analysis made over the average value, and the Bland Altman standard deviation Graf was evaluated whether the measurements were within the safety ranges.

The reliability of this method, which is also quite common in our country, and the determination of the differences between the practitioners have become important, therefore, the reliability of the application of the Graf US method has been tried to be evaluated.

In this context, when the results of the study were evaluated, the intraobserver difference in alpha angles in the measurements made by the researchers participating in the study was realized at the level of maximum 4 at least 1 degree when the Bland Altman table groups were examined and the majority of the measurements in all tables Bland Altman remained within the safe margin of +1.96 to -1.96. The mean deviation of the measurement differences in beta angles spread over a wider band than the alpha values, but most of the measurements remained within the safe margin of Bland Altman +1.96 to -1.96. The differences in beta angles can be partially ignored due to the fact that the alpha angle is more decisive in sonographic evaluation, but if the standard deviation differences in alpha angles spread over a wider band, Graff US typing, which determines the treatment, could also lead to different results. Since large differences in possible alpha angles shall also affect sonographic hip types, there is a risk of adversely affecting treatment and follow-up.

The fact that this study was conducted among experienced researchers who frequently applied the Graf method led to the expectation that the results would be more consistent. However, the differences in the results show that there is a moderate and substantial level of agreement, and the experience of the method shows that there is a risk that there may be less harmony among medical professionals.

In the study, the Cohens Kappa values of the Graf hip typing that the researchers decided morphologically were examined and the compatibility between the researchers was evaluated according to Landis and Koch's Kappa evaluation table, and while the compatibility between some researchers was determined at a moderate level, a significant level of compatibility was observed in others. "However, although there are differences in typing, the effects on treatment recommendations are not different."

Expanding this study in the future to include more inexperienced surgeons and other medical professionals shall carry the risk of decreasing the reliability of the method, but shall allow possible more accurate results and evaluations, thus being important in assessing the reliability of the method.

The fact that the compatibility of the Graf US scanning method between experienced surgeons remains at an acceptable level takes into account the conclusion that it would be more accurate to use the method in screening and follow-up, especially by experienced professionals. Over time, with the increase in the experience of other professionals, the differences in diagnosis and treatment and follow-up shall decrease accordingly.

Conflict of interest
No conflict of interest was declared by the authors.

REFERENCES

1. Graf R (1978) The diagnosis of congenital hip-joint dislocation by the ultrasonic Combound treatment. Arch Orthop Trauma Surg. 97(2):117-33.
2. Terjesen T, Rundén TO, Tangerud A (1989) Ultrasonography and radiography of the hip in infants. Acta Orthop Scand. Dec;60(6):651-60.
3. Harcke HT (1994) Screening newborns for developmental dysplasia of the hip: the role of sonography. AJR Am J Roentgenol. Feb;162(2):395-7.
4. Novick G, Ghelman B, Schneider M (1983) Sonography of the neonatal and infant hip. AJR Am J Roentgenol. Oct;141(4):639-45.
5. Suzuki S, Kasahara Y, Futami T, Ushikubo S, Tsuchiya T (1991) Ultrasonography in congenital dislocation of the hip. Simultaneous imaging of both hips from in front. J Bone Joint Surg Br. Nov;73(6):879-83.
6. Graf R (1984) Classification of hip joint dysplasia by means of sonography. Arch Orthop Trauma Surg. 102(4):248-55.
7. Ortolani M (1976) Congenital hip dysplasia in the light of early and very early diagnosis. Clin Orthop Relat Res. Sep;(119):6-10.
8. Tümer Y, Ömeroğlu H (1997) Prevention of developmental hip dysplasia in Turkey. Acta Orthop Traumatol Turc. 1:176-181.
9. Oğuz T, Ege A, Güngör Ş, Toppare M, Erdemtok N (1996) Ultrasonographic evaluation of 1099 babies with the Graf method. Arthroplasty Arthroscopic Surgery. 7:64-6.
10. Karapinar L, Sürenkök F, Oztürk H, Us MR, Yurdakul L (2002) The importance of known risk factors in the diagnosis of developmental dysplasia of the hip: Selective ultrasonographic screening. Acta Orthop Traumatol Turc. 36:106-10
11. Dogruel H, Atalar H, Yavuz OY, Sayli U (2008) Clinical examination versus ultrasonography in detecting developmental dysplasia of the hip. Int Orthop. Jun;32(3):415-9.
12. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. Feb 8;1(8476):307-10.
13. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics. Mar;33(1):159-74.
14. Simon EA, Saur F, Buerge M, Glaab R, Roos M, Kohler G (2004) Inter-observer agreement of ultrasonographic measurement of alpha and beta angles and the final type classification based on the Graf method. Swiss Med Wkly. Nov 13;134(45-46):671-7.
15. Bar-On E, Meyer S, Harari G, Porat S (1998) Ultrasonography of the hip in developmental hip dysplasia. J Bone Joint Surg Br. Mar;80(2):321-4.
16. Omeroğlu H, Biçimoğlu A, Koparal S, Seber S (2001) Assessment of variations in the measurement of hip ultrasonography by the Graf method in developmental dysplasia of the hip. J Pediatr Orthop B. Apr;10(2):89-95.