THE INTERNATIONAL SOCIETY FOR GYNECOLOGIC ENDOSCOPY
Promoting MIGS
From Vaginal To Robotics and Beyond
ISGE

theTrocar
Official Online Journal of ISGE

# The Reproducibility of the Endometriosis Fertility Index: Inter- and Intra-Observer Variation in the Least Functional Score

Author: Annelize Barnard [1], Viju Thomas[1]

Affiliation: [1] University of Stellenbosch, Tygerberg Hospital, Department of Obstetrics and Gynaecology, Cape Town, South Africa

## Abstract

**Introduction:** To evaluate intra- and inter-observer agreement in the Least Functional Score (LFS) component of the Endometriosis Fertility Index (EFI) amongst gynecologists. As a secondary outcome, we aimed to stratify results according to reviewer expertise.

**Design:** Prospective study (Canadian Task Force II-1).

**Setting:** A university hospital, two referral hospitals and two private sector clinics.

**Method:** Laparoscopic footage of 20 surgical procedures was recorded and presented to 20 gynecologists: 9 sub-specialists in infertility or endometriosis and 11 general gynecologists. Each reviewer was asked to watch and score all 20 videos using the Least Functional component of the Endometriosis Fertility Index on two occasions, more than a year apart.

**Measurements and Main Results:** Interclass correlation coefficient (ICC) and Weighted Kappa values were used to determine inter-observer agreement. Inter-observer agreement within our cohort was found to be moderate (ICC 0.5; κ 0.485) for the Least Functional score. This was true for the group as a whole as well as the sub-groups with an ICC of 0.53 vs 0.58 and κ 0.520 vs κ 0.565 for the sub- specialists and generalists respectively. While we observed a trend towards higher levels of agreement amongst the sub-specialist group for the individual structures, this did not reach statistical significance. With the exception of a single generalist, the observers in both groups achieved substantial intra-observer agreement.

**Conclusion:** This study found moderate inter-observer agreement with regards to the Least Functional Score component of the Endometriosis Fertility Index and substantial intra-observer agreement for the majority of reviewers who took part in the follow-up study. We conclude that gynecologists of varying levels of expertise are equally capable of using the Endometriosis Fertility Index.

**Key words:** Endometriosis, Laparoscopy, Reproductive Surgery, Benign Gynecology, Assisted reproduction (IVF, ICSI, IUI)

*Corresponding author: Annelize Barnard*

*DOI: 10.36205/trocar5.2024018*
*Received:14.10.2024– Accepted 10.11.2024*

## Introduction:

Endometriosis is a major contributor of gynecological morbidity, specifically with relation to sub-fertility and pelvic pain. It is therefore vital that we are able to scientifically explore this field. This necessitates collaborative efforts between clinicians and centers that may be separated, not only geographically, but also by terminology. Multicenter- and meta-analytical studies are dependent on our ability to uniformly classify disease and compare similar patients and stages in terms of management and outcomes. Currently the study of endometriosis is hampered by the lack of a descriptive classification system that is both scientifically sound and of prognostic value – with specific reference to fertility outcomes.

Most of the endometriosis classification systems historically and currently in use, have been aimed at predicting the likelihood of pregnancy for a given stage of the disease. However, most classification systems are not sufficiently predictive to be useful in clinical practice, as poor correlation is found between the extent of disease as classified and pregnancy outcomes in the older models including the American Fertility Society (AFS) score and its successor, the revised American Society of Reproductive Medicine (rASRM) score (1–3).

To address this, Adamson and Pasta collected prospective data on 801 consecutively diagnosed and treated patients and used regression analysis to derive the factors most predictive of pregnancy (4–6). These factors were then used to develop a system known as the Endometriosis fertility index (EFI). The Endometriosis Fertility Index is the first and only classification system that was developed using regression analysis of prospective data and has been proven predictive of fertility outcomes (4–8). It has been widely accepted by the academic community as illustrated by its incorporation into the World Endometriosis Research Foundation (WERF) Endometriosis Phenome and Biobanking Harmonization Project (EPHect) standard surgical form (SSF) as well as World Endometriosis Society Toolbox for surgical staging of endometriosis (9,10).

The EFI is intended for use with infertility patients who are surgically staged and assumes the presence of normal gametes and a normal uterus. It consists of historical factors as well as surgical factors. Surgical factors include the AFS score as well as the Least Functional score (LFS). The LFS is determined at surgery by evaluating and scoring the fallopian tube and ovaries. The

LFS component of the EFI has been shown to be the most predictive of fertility outcome and thought to reflect the function of the reproductive organs (5–7).

As of date, three independent investigators have assessed the EFI, all of whom found the EFI predictive of pregnancy rates (7,8,11). Adamson and Pasta did perform a sensitivity analysis to determine the effect of potential inter-observer differences in the LFS on the EFI during their original research (4). This was however based on statistical assumptions regarding the expected distribution of variability and has not been empirically tested. The historical component consists of objective factors and the AFS has previously been investigated in terms of reproducibility. The aim of this study was therefore restricted to the evaluation of inter-observer agreement in the scoring of the LFS. We also stratify our results by reviewer expertise.

## Materials and methods:

The protocol for this study was approved by the Health Research Ethics Committee of the University of Stellenbosch prior to commencement (Ethics reference number: 14/02/041). The primary outcome was inter-observer agreement as measured by ICC and weighted Kappa scores. As a secondary outcome, the reviewers were stratified by expertise and the agreement within the two groups compared in terms of inter- and intra-observer variability.

### *Study-population:*

The study involved a sample of patients and a selection of clinicians. The latter consists of two groups: general gynecologists and endometriosis/infertility specialists. The groups are delineated below. The sample sizes were determined after consultation with a medical statistician. It must be noted that power calculations for inter-rater studies require knowledge or assumption of the expected ICC. As the ICC for the EFI was not known (hence the need for this study), these calculations proved difficult. As a limited number of endometriosis/infertility specialists practice in the greater Cape Town area, we opted to approach them and match the willing participants with an equal number of general specialists. 20 videos, reviewed by 30 reviewers was calculated to adequately power the study (80%) to detect an effect of 0.2 in the Intra-class Correlation coefficient (ICC) (ICC 0.5-0.7).

### *Patients:*

The clinical notes of patients booked for elective laparoscopic gynecological surgery at Tygerberg Hospital in Cape Town, South Africa were reviewed pre-operatively to identify patients suitable for inclusion into the study. On admission, these patients were counselled regarding the study and written consent obtained for inclusion. Patients were managed by their treating clinician and evaluated for surgery independently. The choice of operative route and surgical intervention lay with the treating clinician and was not influenced by the study.

### *Video-Footage:*

During the surgery of the selected patients, the video output was recorded digitally. This was done at the end of the procedure when the pelvic structures were reviewed.

The footage was assessed by the primary author and deemed suitable if the visual quality was good and the adnexa clearly visualized. From the suitable footage, 20 videos representative of the spectrum of adnexal disease were selected.

### *Reviewers:*

Reviewers were selected from gynecologists active at multiple centers affiliated with Tygerberg Hospital, Cape Town South Africa. This included gynecologists active in the public health sector (Tygerberg hospital - which is an academic referral center, and two of its referring hospitals) as well as gynecologists from 2 clinics in the private health sector. Five centers were therefore involved. The reviewing gynecologists consisted of two groups namely generalists and sub- specialists. Sub-specialists: Individuals who are considered experienced in endometriosis and/or infertility surgery were identified. These individuals were qualified gynecologists who routinely perform laparoscopic surgery and are either advanced endoscopic surgeon (regularly performing excisional surgery for r-ASRM stage 3 and 4 endometriosis) or sub-specialists in reproductive medicine and therefore qualified in endometriosis surgery. 10 such specialists were identified and invited to join the study, of whom 9 enrolled and 8 completed the review process. 1 clinician did not respond and 1 enrolled but did not return any forms.

### *Generalists:*

General specialists affiliated with Tygerberg Hospital were eligible. These were qualified gynecologists who perform laparoscopic surgery but do not perform excisional surgery for r-ASRM stage 3 and 4 endometrioses. Sub-specialists were excluded from this group. 15 general specialists were identified and invited to join the study, of whom 11 enrolled and 10 completed the review process. 4 clinicians did not respond. The reviewers in both groups were counselled regarding the purpose of the study and written consent was obtained for inclusion in the study. All reviewers were given a digital copy of the study videos and a scoring sheet for each video. Instructions on how to perform the scoring were included on the scoring sheet. In order to avoid bias no further instruction on how to perform the scoring was provided by the researcher. After one year, participants were asked to repeat the process. 7 clinicians (2 Sub-specialists and 5 general specialists) completed the second round within the allotted timeframe.

### Data management and statistical analysis:

The data was analyzed with the help of a statistician from the Biostatistics unit, Centre for Evidence Based Health Care, Faculty of Health Sciences, University of Stellenbosch. Stata Version 13.1 was used. Interclass correlation coefficient (ICC) values were used to calculate the level of agreement within each group as well as agreement overall. We also represent our findings on intra- observer agreement by Kappa values for absolute agreement as well as weighted Kappa values for the weighted agreement.

### Results:

A total of 20 clinicians were recruited in the first round. 1 reviewer submitted incomplete forms and 1 did not return any forms. These reviewers were excluded from the study. This brought the total to 18 reviewers: 8 in the sub-specialist group and 10 in the generalist group. Two sub-specialists and 5 generalists took part in the second assessment. Our findings are presented in Table 1 for weighted and absolute inter-observer agreement on the LFS.

| Least Functional Score | | | | |
|---|---|---|---|---|
| | Absolute Agreement | | Weighted Agreement | |
| | κ - Value | 95% CI | κ - Value | 95% CI |
| Overall | 0.148 | 0.107-0.213 | 0.485 | 0.351-0.651 |
| | | | | |
| Sub-specialists | 0.214 | 0.157-0.290 | 0.520 | 0.339-0.732 |
| General Specialists | 0.143 | 0.099-0.216 | 0.565 | 0.417-0.753 |

Table 1 - Absolute and Weighted Inter-observer Agreement – Least Functional Score

Table 2 presents the results per anatomical structure. Table 3 summarises the results as ICC values. 95% confidence intervals are indicated. Tables 4 and 5 contains the agreement measures for the interpretation of Kappa values as per Landis and Koch and ICC values as per Koo and Li (12,13). In the first round, we found the overall inter-observer agreement for the Least Functional Score to be moderate (ICC 0.5). Both groups achieved moderate agreement for almost all structures by ICC. However, overall absolute agreement on the exact score along the nominal scale was, as expected, slight (κ = .148). For the second assessment, we found the overall inter-observer agreement for the Least Functional Score to be moderate (ICC 0.45). Both groups achieved moderate agreement for almost all structures by ICC. However, overall absolute agreement on the exact score along the nominal scale was again, as expected, slight (κ = .121).

Table 2 - Summary of Results per structure – Weighted Inter-observer Agreement

| **Weighted Inter-observer Agreement** | | | | | | |
|---|---|---|---|---|---|---|
| | Overall | | Subspecialists | | Generalists | |
| | Kappa | CI | Kappa | CI | Kappa | CI |
| Left Fallopian Tube | 0.464 | 0.305-0.661 | 0.501 | 0.277-0.773 | 0.474 | 0.319-0.666 |
| | | | | | | |
| Right Fallopian Tube | 0.614 | 0.460-0.812 | 0.706 | 0.513-0.939 | 0.579 | 0.425-0.774 |
| Left Fimbria | 0.529 | 0.353-0.747 | 0.612 | 0.363-0.919 | 0.51 | 0.368-0.687 |
| Right Fimbria | 0.578 | 0.389-0.813 | 0.689 | 0.483-0.944 | 0.513 | 0.324-0.744 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Left Ovary | 0.534 | 0.270-0.889 | 0.541 | 0.235-0.941 | 0.52 | 0.276-0.853 |
| Right Ovary | 0.572 | 0.350-0.872 | 0.59 | 0.347-0.897 | 0.543 | 0.305-0.868 |
| Left LF Subtotal | 0.488 | 0.333-0.676 | 0.55 | 0.364-0.771 | 0.534 | 0.411-0.687 |
| Right LF Subtotal | 0.599 | 0.476-0.753 | 0.625 | 0.422-0.860 | 0.633 | 0.496-0.804 |
| LF Score | 0.485 | 0.351-0.651 | 0.52 | 0.339-0.732 | 0.565 | 0.417-0.753 |

Table 3 - Summary of results per structure – Interclass correlation coefficient (ICC)

|  | Round 1 |  | Round 2 |  | Round 1 |  | Round 2 |  | Round 1 |  | Round 2 |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | Overall | 95% CI | Overall | 95% CI | Sub-specialists | 95% CI | Sub-specialists | 95% CI | Generalist | 95% CI | Generalist | 95% CI |
| Left Tube | 0,51 | (0.35-0.71) | 0,41 | (0,12-0,64) | 0,56 | (0.37-0.76) | 0,73 | (0,44-0,88) | 0,49 | (0.32-0.69) | 0,33 | (0,12-0,59) |
| Right Tube | 0,63 | (0.48-0.79) | 0,44 | (0,24-0,66) | 0,72 | (0.57-0.85) | 0,71 | (0,38-0,88) | 0,59 | (0.43-0.77) | 0,35 | (0,15-0,59) |
| Left Fimbria | 0,55 | (0.39-0.74) | 0,50 | (0,30-0,71) | 0,68 | (0.50-0.84) | 0,73 | (0,43-0,88) | 0,52 | (0.36-0.72) | 0,42 | (0,20-0,65) |
| Right Fimbria | 0,61 | (0.46-0.78) | 0,46 | (0,27-0,67) | 0,76 | (0.59-0.89) | 0,81 | (0,58-0,92) | 0,53 | (0.35-0.72) | 0,37 | (0,17-0,61) |
| Left Ovary | 0,55 | (0.39-0.74 | 0,33 | (0,15-0,57) | 0,56 | (0.38-0.75) | 0,59 | (0,22-0,81) | 0,53 | (0.36-0.72) | 0,24 | (0,06-0,50) |
| Right Ovary | 0,59 | (0.43-0.77) | 0,37 | (0,18-0,61) | 0,60 | (0.42-0.78) | 0,68 | (0,35-0,86) | 0,56 | (0.39-0.74) | 0,27 | (0,07-0,52) |
| Left LFS Subtotal | 0,50 | (0.35-0.69) | 0,48 | (0,30-0,69) | 0,56 | (0.39-0.75) | 0,64 | (0,30-0,84) | 0,55 | (0.38-0.73) | 0,40 | (0,20-0,64) |
| Right LFS subtotal | 0,61 | (0.46-0.77) | 0,47 | (0,29-0,68) | 0,64 | (0.47-0.80) | 0,93 | (0,84-0,97) | 0,65 | (0.49-0.80) | 0,35 | (0,16-0,59) |
| LFS | 0,50 | (0.35-0.69) | 0,46 | (0,27-0,67) | 0,53 | (0.36-0.73) | 0,85 | (0,66-0,94) | 0,58 | (0.41-0.76) | 0,38 | (0,17-0,62) |

Table 4 - Interpretation of Kappa Values as per Landis and Koch (12)

| Kappa Score | Interpretation |
|---|---|
| <1 | No agreement |
| 0.0-0.2 | Slight agreement |
| 0.21-0,4 | Fair agreement |
| 0.41-0.6 | Moderate agreement |
| 0.61-0.8 | Substantial agreement |
| 0.81-1.0 | Almost agreement |

Table 5 - Interpretation of ICC values as per Koo and Li (13)

| ICC | Reliability |
|---|---|
| <0.5 | Poor |
| 0.5-0.75 | Moderate |
| 0.75-0.9 | Good |
| >0.9 | Excellent |

With regards to the performance of the sub-specialists versus generalists; no statistically significant difference was detected for the LFS or for any of the adnexal structures. There is however a clear trend visible, with the sub-specialists consistently achieving higher Kappa values for all the structures.

The generalists achieved slightly better agreement in the right LFS subtotal as well as in the Least Functional score, but this did not reach statistical significance. Apart from one generalist, all reviewers achieved substantial intra-observer agreement as indicated in Table 6.

Table 6 Intra-observer reliability

| | | | CI (95%) | |
|---|---|---|---|---|
| | Reviewer | Kappa | Lower | Upper |
| Subspecialists | A | 0,81 | 0,68 | 0,94 |
| | B | 0,71 | 0,37 | 1,05 |
| Generalists | C | 0,81 | 0,68 | 0,93 |
| | D | 0,62 | 0,35 | 0,89 |
| | E | 0,76 | 0,58 | 0,94 |
| | F | 0,14 | 0,00 | 0,29 |
| | G | 0,89 | 0,82 | 0,96 |

## Discussion:

Our aim with this study was to assess the reproducibility of the Least Functional component of the Endometriosis Fertility Index and as a secondary outcome, to stratify this by level of expertise. We found that weighted agreement was moderate (ICC 0.5; κ = .485) with no statistically significant difference in the performance of sub-specialists versus generalists. We only evaluated the LFS and not the complete EFI as the other components of the EFI are either objective, such as the historic factors, or previously studied, such as the AFS.

While the EFI differs from the r-ASRM in its design and aim, it is perhaps useful to compare the reproducibility of the two systems. The latter is widely used in the evaluation of infertility patients to assess structural damage that may impact on fertility, despite not having been validated for this purpose. If the EFI is to challenge this status quo, it needs to be reproducible, in addition to being predictive of fecundity. Our findings on inter-observer agreement are similar to those of Schliep and colleagues' study from the 'Endometriosis: Natural History, Diagnosis and Outcomes study' group. They found moderate inter-observer reliability (κ = .44) for the r-ASRM comparable to our findings of moderate agreement (κ = .485) for the LFS (14). Schliep and colleagues found academic experts to be more reliable for the diagnosis of disease than the other experts. (κ = .79 vs. κ = .58) (14). We found no statistically significant difference between the generalist and sub-specialist groups. It must however be noted that our outcomes and the composition of our groups differed from theirs and our finding can therefore not be compared directly.

Our results on inter-observer agreement amongst the groups are in keeping with those of Buchweitz and colleagues who found no difference in the accuracy of specialists versus trainees in their study on the staging of endometriosis using the r-ARSM score (15). Our findings on agreement are also comparable to theirs in that they also found only marginal inter-observer correlation (Kendall coefficient of .14) It must however be noted that their study included no sub-specialists, therefore our groups were dissimilar to theirs and the results cannot be directly compared (15). With regards to intra-observer agreement, all but one reviewer achieved substantial or almost perfect agreement with kappa values between 0.62 and 0.89.

## Conclusion:

We found the LFS component of the EFI to be moderately reproducible with no statistically significant difference in the performance of sub-specialists and general gynecologists. We therefore conclude that the EFI can be used in clinical practice by clinicians of varied levels of experience. Our findings may aid others in planning adequately powered studies involving the EFI.

## References:

1. Marana R, Rizzi M, Muzii L, Catalano GF, Caruana P, Mancuso S, et al. Correlation between the American Fertility Society classifications of adnexal adhesions and distal tubal occlusion, salpingoscopy, and reproductive outcome in tubal surgery. Fertil Steril. 1995 Nov;64(5):924–9.

2. Guzick DS, Silliman NP, Adamson GD, Buttram VCJ, Canis M, Malinak LR, et al. Prediction of pregnancy in infertile women based on the American Society for Reproductive Medicine's revised classification of endometriosis. Fertil Steril. 1997 May;67(5):822–9.

3. Vercellini P, Fedele L, Aimi G, De Giorgi O, Consonni D, Crosignani PG. Reproductive performance, pain recurrence and disease relapse after conservative surgical treatment for endometriosis: The predictive value of the current classification system. Hum Reprod. 2006;21(10):2679–85.

4. Adamson GD, Pasta DJ. Endometriosis fertility index: the new, validated endometriosis staging system. Fertil Steril. 2010 Oct;94(5):1609–15.

5. Adamson GD. Endometriosis classification: an update. Curr Opin Obstet Gynecol. 2011 Aug;23(4):213–20.

6. Adamson GD. Endometriosis Fertility Index: is it better than the present staging systems? Curr Opin Obstet Gynecol. 2013 Jun;25(3):186–92.

7. Tomassetti C, Geysenbergh B, Meuleman C, Timmerman D, Fieuws S, D'Hooghe T. External validation of the endometriosis fertility index (EFI) staging system for predicting non-ART pregnancy after endometriosis surgery. Hum Reprod. 2013 May;28(5):1280–8.

8. Wei D, Yu Q, Sun A, Tian Q, Chen R, Deng C, et al. [Relationship between endometriosis fertility index and pregnancies after laparoscopic surgery in endometriosis-associated infertility]. Zhonghua Fu Chan Ke Za Zhi [Internet]. 2011 Nov [cited 2018 Feb 1];46(11):806–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22333226

9. Becker CM, Laufer MR, Stratton P, Hummelshoj L, Missmer SA, Zondervan KT, et al. World Endometriosis Research Foundation Endometriosis Phenome and Biobanking Harmonisation Project: I. Surgical phenotype data collection in endometriosis research. Fertil Steril. 2014 Nov;102(5):1213–22.

10. Johnson NP, Hummelshoj L, Adamson GD, Keckstein JJ, Taylor HS, Abrao MS, et al. World Endometriosis Society consensus on the classification of endometriosis. Hum Reprod. 2016 Dec;1–10.

11. Wang W, Li R, Fang T, Huang L, Ouyang N, Wang L, et al. Endometriosis fertility index score maybe more accurate for predicting the outcomes of in vitro fertilisation than r-AFS classification in omen with endometriosis. Reprod Biol Endocrinol. 2013;11(1).

12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.

13. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med [Internet]. 2016 Jun;15(2):155–63.

14.S chliep KC, Stanford JB, Chen Z, Zhang B, Dorais JK, Boiman Johnstone E, et al. Interrater and intrarater reliability in the diagnosis and staging of endometriosis. Obstet Gynecol. 2012 Jul;120(1):104–12.

15. Buchweitz O, Wulfing P, Malik E. Interobserver variability in the diagnosis of minimal and mild endometriosis. Eur J Obstet Gynecol Reprod Biol. 2005 Oct;122(2):213–7.