Intraobserver and interobserver reliability of videotaped laparoscopy evaluations for endometriosis and adhesions

Philomeen Weijenborg, Moniek ter Kuile and Frank Willem Jansen.
ABSTRACT

Objective: To determine the intra- and interobserver reliability of evaluations during videotaped laparoscopy, with real-time laparoscopy as the “gold standard.”

Design: Prospective evaluation.

Setting: University hospital.

Patients: Women who underwent laparoscopy for chronic pelvic pain, sterilization, or infertility workup.

Intervention: Real-time laparoscopies were videotaped and scored, then later reassessed.

Main Outcome Measure: Intra- and interobserver levels of agreement between evaluations for endometriosis and adhesions.

Results: With the use of reassessments on 90 (videotaped) laparoscopies, the intra- and interobserver levels of agreement between the scorings for endometriosis were found to be substantial, except for ovarian implantations. A high agreement was found in the staging of endometriotic disease. The intra- and interobserver levels of agreement for scoring adhesions were only fair to moderate, and a substantial number of differences between measurements in adhesion total scores was found. No systematic difference between the number of disagreements was observed in either setting.

Conclusions: Although special attention has to be given to the assessments of ovarian lesions, the evaluations of videotaped laparoscopies for endometriosis were reliable and justified the use of recorded findings. Because evaluations of adhesions during videotaped laparoscopy are not reliable, in some cases a second laparoscopy may need to be performed.
Introduction

Performing laparoscopy is a common tool for gynecologists in the case of diagnostic or therapeutic procedures. To record the findings, videotaping of this procedure has been introduced. Gradually, videotaped laparoscopies found general acceptance for residential training, informing the patient of the findings, and requesting second opinions. Video recordings have also been introduced as evidence both for and against the operator in medical malpractice proceedings [Corson 1995]. In these circumstances, it is a prerequisite that evaluations of videotaped laparoscopies are consistent with real-time laparoscopic findings, the so-called “gold standard.”

In a study by Bowman [Bowman 1995], scorings of adnexal adhesions during real-time laparoscopy in women who had been diagnosed previously with adhesions were compared with assessments of videotaped laparoscopies by 2 separate assessors. A large variation in adhesion scorings between 2 of the 3 observers and a poor level of agreement on subdivisions of adhesion total scores were reported.

In another study by Corson [Corson 1995], comparisons were made between scorings by 1 operating surgeon during real-time laparoscopies of women who were diagnosed with adhesions and reassessments of videotaped laparoscopies by 4 separate observers, before adhesiolysis, at second look, and 4 months later. An acceptable intra- and interobserver variability in scoring laparoscopic diagnosis of pelvic adhesions was found.

The reproducibility of the revised American Fertility Society classification for endometriosis was evaluated [Hornstein 1993]. When 5 assessors reviewed videotaped laparoscopies of patients with endometriosis twice, an acceptable agreement between the peritoneal scores was observed. However, great variability in the ovarian endometriosis and cul-de-sac components of the classification was also found.

Rock [Rock 1995] found a fair level of agreement between assignments of the endometriotic disease stage by 22 surgeons who scored real-time laparoscopies of women with endometriosis and 1 blinded assessor who reviewed visual documentation.

Recently, Buchweitz [Buchweitz 2005] evaluated the interobserver variability in the diagnosis of minimal and mild endometriosis. A digital videotape of 3 patients (1 patient with stage I endometriosis, 1 patient with stage II endometriosis and 1 patient without endometriosis) was presented to 108 gynecologists. In this study, the number and location of the endometriotic lesions varied substantially between the observers, and a correct classification of the endometriotic disease stages I and II was found in only 22% and 13% of the cases, respectively.

All of these reliability studies, with the exception of the last one, were performed in selected populations (i.e., women previously diagnosed with adhesions or endometriosis). Further, in most studies, assessors were informed of the clinical history, complaints, and/or treatment of the patients. Although it is acknowledged that the accuracy of a diagnostic test increases with increased prevalence of the target condition and with the provision of
clinical information [Whiting 2004], no studies have been conducted in a more general population, with totally “blinded” assessors.

Therefore, the main purpose of the present study was to investigate intra- and interobserver reliability of evaluations by assessors, who viewed videotaped laparoscopies compared with real-time laparoscopies in a sample of a heterogeneous population of women with endometriosis and/or adhesions or without disease. Clinical information would be available only for the operating surgeon and not for the assessors, who also were blinded to one another’s measurements. In this way, we expected to determine the level of agreement between measurements in both the intra- and interobserver settings of scoring the presence or absence of endometriosis and adhesions and scoring the severity and extent of the disease.

**Material and Methods**

Consecutive women from a gynecologic outpatient clinic of a university hospital for whom a diagnostic laparoscopy was indicated for chronic pelvic pain (CPP), sterilization, or infertility workup (FER) were invited to participate in the study.

The procedure of the laparoscopy was standardized. A 2-trocar double puncture technique was used. All laparoscopies were performed with the use of general anesthesia in the same operating theatre, and all procedures were videotaped. The operating surgeons and the team of nurses were instructed to use the same standardized procedure regarding which specific structures were to be recorded to obtain a detailed view of the pelvis and abdominal cavity. Recording had to be finished before any surgical intervention was started. In this way, only the diagnostic part of the laparoscopy procedure was videotaped.

The findings on viewing real-time and videotaped laparoscopies were both scored on a sheet that had been designed especially for this study.

**Endometriosis**
The assessor had to mark whether endometriosis was present, absent, or not to be determined. The scoring sheet of the revised classification of endometriosis of the American Fertility Society was used to assess the severity and location of endometriosis, which resulted in a total score (range 0-60). The stages of the disease that were calculated from the total score are defined as Stage I (minimal; score 1-5), Stage II (mild; score 6-15), Stage III (moderate; score 16-40) and Stage IV (severe; score, >40) [Revised AFS Classification 1985].

**Adhesions**
The assessor had to mark whether adhesions were present, absent, or not to be determined. A scoring form was used to assess the severity (16 sites) and extent (11 sites) of the
adhesions, both pelvic and abdominal [Mage 2000]. The total adhesion score (range 0-97) was calculated as described by the Adhesion Scoring Group [Adhesion Scoring Group 1994].

**Quality of the Videotaped Laparoscopy**
To score the quality of the videotape, a visual analogue scale of 10 cm was used (range 0 = very bad to 10 = excellent).

The Institutional Review Board of the Leiden University Medical Center approved this prospectively designed evaluation of scorings by gynecologists during real-time and videotaped laparoscopies. All patients who participated in this study gave their informed consent.

**Evaluation of Videotaped Laparoscopies**
To determine the intraobserver reliability of the videotaped evaluations, the level of agreement between measurements by the same surgeon during real-time and videotaped laparoscopy was obtained. To determine the interobserver reliability, the level of agreement between measurements by 2 different observers who scored the same videotaped laparoscopy was used.

Each operating surgeon assessed a sample of his own videotaped laparoscopies and those of his colleagues. Appointments to look at the videotapes were made by a research assistant. Each tape was viewed in total. On request, the observer was permitted to review a portion of the tape for clarification. To prevent fatigue, scores were made for a maximum of 1 hour or 10 videotapes per session. During video assessment, the gynecologists and the research assistant were unaware of the clinical history of the patient, the indication for the operation, or the name and scores of the original operating surgeon. We deliberately made the choice to ask only surgeons from the same hospital and department to participate in the study to avoid a source of bias that could be caused by differences in expertise and policy of every day practice between medical centers.

A sample was taken at random from the original videotaped laparoscopies, with the use of the randomization function of the Statistical Package for Social Sciences (version 11.0; SPSS Inc., Chicago, IL). The distribution of women who underwent laparoscopy because of chronic pelvic pain, sterilization, or FER in the original sample was preserved. The number of laparoscopies that were performed by each operating surgeon varied between 4 and 35 procedures. In the event that the operating surgeon had made >10 videotapes, a random sample of 10 was drawn from his own tapes. Surgeons who made <10 videotapes kept his/her own videotapes to assess; the total amount of videotapes was upgraded to 10 by taking, at random, videotapes from those surgeons, who had originally made >10 tapes. To obtain a total number of 20 videotapes for assessment, the remaining 10 tapes were allocated at random to each assessor. The sequence of the videotapes that were presented for scoring was also at random.
**Statistical Analysis**

Kappa (κ) statistics were used to determine the level of agreement between 2 measurements or observers on a categoric scale (i.e., the finding [yes or no] for endometriosis, adhesions, and the stages of endometriotic disease). With κ statistics, the amount of agreement between a pair of observations, over and above what is expected by chance alone, is calculated. When κ equals 1, perfect agreement is implied; whereas when κ equals 0, the agreement is no better than that which would be obtained by chance. Landis and Koch [Landis 1977] have given an indication of judging intermediate values. For most purposes κ ≤ 0.20 represents poor agreement; values between 0.21 ≤ κ ≤ 0.40 represent fair agreement; values between 0.41 ≤ κ ≤ 0.60 represent moderate agreement; values between 0.61 ≤ κ ≤ 0.80 represent substantial agreement, and a value of κ > 0.80 indicates good agreement. McNemar tests were used to estimate the probability of a systematic difference between the number of disagreements.

To determine the level of agreement between measurements on a continuous scale (such as in case of total scores for endometriosis and adhesions), Bland Altman plots [Bland 1986; Khan 2001] were constructed. The difference against the mean of the measurements for each subject in the study is used. If the average difference is 0, no bias in results is inferred, which implies that on average the duplicate readings agree.

The British Standards Institution repeatability coefficient, by definition 2 times the SD of the mean of the differences, indicates the maximum difference that is likely to occur between 2 measurements. It provides a measure of agreement and can be used as a comparative tool. The range that encompasses 95% of the differences between measurements (d ± 2 SD) is limited by the so-called upper and lower limits of agreement.

Finally, with t-tests for independent samples, the effect of the quality of the videotapes on the scores for endometriosis and adhesions was evaluated.

**Results**

One hundred fifty-one laparoscopies were performed and recorded on videotape by or under the supervision of 9 senior gynecologists of the Department of Gynecology, Leiden University Medical Center. For 11 of the total number of tapes (7%), the name of the gynecologic supervisor was not indicated in the records. Only reassessments of videotaped laparoscopies that had been scored by gynecologic staff members as operating surgeon or supervisor could be used; therefore, these 11 tapes had to be excluded. From the remaining 140 laparoscopies, a final research sample of 90 (videotaped) laparoscopies was constructed.

In this sample, endometriosis was found only in the FER group, in 53% of the cases. Adhesions were seen in 66% of the cases in the chronic pelvic pain group and in approximately
one half of the cases in the sterilization and FER group. In 10 cases of the FER group (19%), adhesions and endometriosis were found. Twenty-four laparoscopies (27% of all cases) did not show any endometriosis or adhesions.

The time span between the last recording and the first assessment was at least 8 months. Thorough viewing of 1 tape took an average of 5 minutes (range 4-7 minutes). In total 56 (videotaped) laparoscopies could be used to determine the intraobserver reliability; 90 videotapes could be used to obtain the interobserver reliability. However, on analysis, the total numbers turned out to be less because in 8 of the 180 cases, reassessments were impossible. Diverse reasons were mentioned (for instance, the videotape was too dark, the view was hampered by adhesions, or the observer was unable to give a scoring because palpation of the lesion was not possible). In only 1 case could a videotape not be assessed because of technical problems.

**Intraobserver Reliability**

As illustrated in Table 1, the level of agreement between the scorings (present or absent) for endometriosis that were made by the operating surgeon during real-time laparoscopy and videotaped laparoscopy was substantial ($\kappa = 0.75$), whereas for adhesions the level of agreement was fair ($\kappa = 0.38$).

McNemars tests indicated no systematic difference between the number of disagreements, which indicated that disagreements were distributed equally among the evaluations made during real-time and videotaped laparoscopy. For scoring endometriosis, however, a trend towards a systematic difference in disagreements was found ($p = .06$) in the direction of the videotaped laparoscopy. The operating surgeon who viewed his videotaped laparoscopy was more often inclined systematically to score the presence of endometriosis in cases on which he scored “no endometriosis” during real-time laparoscopy.

**Table 1**  *Intraobserver level of agreement for the presence or absence of endometriosis and adhesions*

<table>
<thead>
<tr>
<th>Real-time laparoscopy</th>
<th>Videotaped laparoscopy</th>
<th>Kappa ($\kappa$)</th>
<th>95% CI</th>
<th>McNemar ($p$-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endometriosis (52)*</td>
<td>Yes</td>
<td>0.75</td>
<td>0.55-0.95</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>5</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>Adhesions (55)*</td>
<td>Yes</td>
<td>0.38</td>
<td>0.13-0.63</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>9</td>
<td>17</td>
<td></td>
</tr>
</tbody>
</table>

Values are number of videotapes. * in parenthesis: total number of videotapes.
In 87% of the repeated measurements, the surgeons agreed on the stage of endometriotic disease (Table 2). In 5 cases (10%), a difference of 1 stage was observed. In 2 cases (3%), a difference of 2 stages occurred that was caused by a marked difference in endometriosis total scores. Both surgeons who viewed the videotaped laparoscopy scored “deep endometriosis in the ovary,” which corresponded with a total score of 16 or 20, whereas during real-time laparoscopy they indicated a superficial ovarian lesion corresponding with a score of 4.

The level of agreement between the measurements for the severity and extent of endometriosis and adhesions was indicated in Bland Altman plots (Figs. 1 and 2). Because 0 was found in the 95% confidence interval (CI) of the mean of the differences for endometriosis total scores (score - 0.9; 95% CI - 1.9; 0.6) and for adhesion total scores (score - 0.3; 95% CI - 0.7; 0.01), it was inferred that, between the 2 measurements, no bias had occurred. Surgeons who assessed their own videotaped laparoscopies did not score systematically higher or lower than when they scored their real-time laparoscopies.

The repeatability coefficient for endometriosis total scores was 7.2, whereas for the adhesion total scores it was 4.8. In case of endometrioses total scores for 41 of the 52 cases the differences in total scores counted 0. For 9 cases the counts for the differences were found within the limits of agreement (- 8; 6). The 2 outlying differences in the FER group resulted from the substantial difference in total scores for deep and superficial ovarian endometriosis.

The total adhesion score could be calculated for only 48 of the 56 videotaped laparoscopies (86%). In 2 cases, assessment of the severity and extent of the adhesions was impossible because of insufficient viewing material and doubts about the diagnosis. In 6 other cases, the type and/or extent of the adhesion was not indicated on the scoring form. Twenty-three differences (48%) counted 0, whereas another 23 differences (48%) had counts within the limits of agreement (- 5; 4.5). Both cases in which the limits of agreement were exceeded, 1 case in the chronic pelvic pain and the other case in the FER group, resulted from a lower score during real-time than during videotaped laparoscopy without a clear reason.

Table 2  Intraobserver level of agreement for stage of endometriosis

<table>
<thead>
<tr>
<th>Real-time laparoscopy</th>
<th>Videotaped laparoscopy</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No endometriosis</td>
<td>Stage I</td>
<td>Stage III</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>No endometriosis</td>
<td>36</td>
<td>5</td>
<td>2</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>9</td>
<td>2</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
<td>14</td>
<td>2</td>
<td>52</td>
<td></td>
</tr>
</tbody>
</table>

Values are number of videotapes.
Interobserver Reliability

As can be seen in Table 3, the level of agreement between the scorings (present or absent) for endometriosis by raters A and B was substantial ($\kappa = 0.75$), whereas for adhesions the level of agreement was moderate ($\kappa = 0.55$). As McNemars tests indicated, no systematic differences between the number of disagreements for the 2 raters were found, which indicated that the disagreements were distributed equally among evaluations by raters A and B.

Table 3  Interobserver level of agreement for the presence or absence of endometriosis and adhesions

<table>
<thead>
<tr>
<th>Rater A</th>
<th>Rater B</th>
<th>Kappa ($\kappa$)</th>
<th>95% CI</th>
<th>McNemar (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endometriosis (83)*</td>
<td>27  4</td>
<td>0.75</td>
<td>0.59-0.89</td>
<td>0.75</td>
</tr>
<tr>
<td>No</td>
<td>6  46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adhesions (88)a</td>
<td>35  13</td>
<td>0.55</td>
<td>0.39-0.72</td>
<td>0.26</td>
</tr>
<tr>
<td>Yes</td>
<td>7  33</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Values are number of videotapes. * in parenthesis: total number of videotapes.
In 80% of the repeated measurements, both raters A and B agreed on the stage of the endometriotic disease that they assigned (Table 4). In 17 cases (20%), a difference of 1 stage was found ($\kappa = 0.59; 95\% \text{ CI} 0.43; 0.75$).

**Table 4 Inter-observer level of agreement for stage of endometriosis**

<table>
<thead>
<tr>
<th>Rater A</th>
<th>No endometriosis</th>
<th>Stage I</th>
<th>Stage II</th>
<th>Stage III</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No endometriosis</td>
<td>46</td>
<td>6</td>
<td></td>
<td></td>
<td>52</td>
</tr>
<tr>
<td>Stage I</td>
<td>6</td>
<td>19</td>
<td>2</td>
<td></td>
<td>27</td>
</tr>
<tr>
<td>Stage II</td>
<td></td>
<td>2</td>
<td>1</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Stage III</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>52</td>
<td>27</td>
<td>2</td>
<td>2</td>
<td>83</td>
</tr>
</tbody>
</table>

Values are number of videotapes.

Bland Altman graphs were constructed for the endometriosis and adhesion total scores (Figs. 3 and 4). Because 0 was found in the 95% CI of the mean of the differences for endometriosis total scores (score 0.25; 95% CI -0.25; 0.75) and for adhesion total scores (score -0.03; 95% CI -0.59; 0.53), it was inferred that no bias was found between the measurements of raters A and B. Raters A did not score systematically higher or lower than raters B when they scored videotaped laparoscopies. The repeatability coefficient for endometriosis and adhesion total scores was 4.6.

In the case of endometriosis total scores, 65 of 83 differences (67%) counted 0, whereas the other 25 differences (30%) were found to be within the limits of agreement (-4.4; 4.9). One substantial difference of 16 was found in the FER group, because rater A scored deep ovarian lesion, whereas rater B did not indicate an ovarian lesion at all.

The total adhesion score could be calculated for only 65 scorings of the 90 videotaped laparoscopies (72%). In 1 case, the severity of the adhesions was not indicated; in 4 other cases, proper assessment of the tape was not possible because of insufficient lighting or complete lack of view. In the other 20 cases, 1 or both assessors did not indicate the extent of the adhesion on the scoring form. The reason for this omission was unclear: either the assessor forgot to fill in the score on the form, or the assessor was not able to give a score. For 39 of the 65 tapes (60%), the difference in total scores was found to be 0. The remaining 25 values of the differences (38%) were found to lie between the limits of agreement (-6.3; 6.8). One outlier was found in the FER group (score -12) that was caused by a substantial difference in the scores of adhesions between rater A and B without a clear reason.
Quality of the Videotape

When surgeons scored their own videotaped laparoscopy, the mean (± SD) of the quality scores was 6.5 ± 2.7. When both observers scored the same videotaped laparoscopy, no significant difference ($t(88) = 0.13; p = 0.9$) was found between the mean of the quality assessment by raters A (mean 6.3 ± 2.6) and that of raters B (mean 6.2 ± 2.6). We observed that the intra- and interobserver level of agreement in scoring endometriosis or adhesions was not related to the quality score of the videotaped laparoscopy as assessed by the observers.

Discussion

In this study we investigated the intra- and interobserver reliability of evaluations based on videotaped laparoscopies, although that of the real-time laparoscopy was considered as the “gold standard.” There were 3 major findings. First, in both the intra- and interobserver settings, a substantial level of agreement was found between the scorings regarding the presence or absence of endometriosis and for the stages of endometriotic disease. Second, we observed a level of agreement of fair to moderate regarding the presence or absence of adhesions in the intra- and interobserver setting, respectively. Last, disagreements between measurements were distributed equally between both settings.
When comparing our findings with those of previous studies [Hornstein 1993; Corson 1995; Bowman 1995; Rock 1995; Buchweitz 2005], it is important to realize that there are many differences in design, sample and measurement. For example, in contrast to others, our aim was to evaluate the intra- and interobserver reliability of the videotaped evaluations, we used a heterogeneous population of women with endometriosis and/or adhesions or without disease, and our assessors were blinded for clinical information and one another’s measurements. Despite these differences, some of the similarities between the pattern of results that were obtained in the present and previous studies are worth mentioning.

Regarding the presence or absence of endometriosis, our findings were partly in line with the results of a recent study by Buchweitz [Buchweitz 2005]. We found that, in 90% of the cases, the assessors who scored a videotaped laparoscopy agreed with their findings during the corresponding real-time laparoscopy. Although Buchweitz did not give an overall agreement, 94% of the assessors in this study agreed at reassessment of 2 videos of patients with endometriosis, whereas approximately one half of the raters disagreed when they saw an endometriotic lesion at the moment they assessed a videotape of a patient without endometriosis. The author suggested that this number of disagreements could be explained by the fact that some observers probably scored the presence of endometriosis to a higher rate compared with “normal conditions,” because the study took place during a workshop on endometriosis. We also found that an assessor who viewed his own videotaped laparoscopy was more often inclined systematically to see endometriosis in cases in which he scored “no endometriosis” during real-time laparoscopy, although in our study clinical information was not provided.

In line with previous studies [Hornstein 1993; Rock 1995], we found a good reproducibility of the total scores for endometriosis and of endometriotic disease stage. Additionally, a large difference in total scores was also observed in our study when raters disagreed on ovarian implantations, which resulted in disagreement on the stage of the disease. In 2 cases of the intraobserver setting, a change from stage I to stage III was observed because the assessors of a videotaped laparoscopy scored deep endometriosis in the ovary, whereas the surgeon indicated a superficial ovarian lesion during real time laparoscopy. In another 22 cases, a change of 1 stage was found. The clinical relevance and implication of these results depend on the main complaint of endometriosis-related pain or infertility. A variety of treatments are recommended, from nonsteroidal anti-inflammatory agents to hormonal treatment and surgical intervention [Kennedy 2005].

Because the level of agreement on scoring the presence or absence of adhesions was just fair to moderate, we were not surprised to find a substantial number of disagreements between the measurements of the adhesion total scores in both the intra- and interobserver setting. In line with results of a previous study [Bowman 1995] in which a poor level of agreement between the subdivided American Fertility Society scores was found, we had to
conclude that obtaining consistency between measurements on adhesion scoring during real-time and videotaped laparoscopy proved to be difficult. Bowman [Bowman 1995] suggested that varying scores between observers could result from the fact that video images would not allow an observer to inspect an organ or pelvic area in detail. Some others [Hornstein 1993; Corson 1995] also proposed that variation between evaluations could be explained partly by the benefit of the surgeon knowing the patient’s history. Although our study was not designed to do research on this subject, our results were not supportive of these suggestions. We found that surgeons assessing their own videotaped laparoscopy did not score systematically higher or lower than when they scored their real-time laparoscopies. Therefore, during real-time laparoscopy scoring, the surgeons seemed not to be biased by knowing patients’ history. These and other factors such as the complexity of the adhesion scoring system could explain the adhesion scoring problem.

The external validity of our results is limited by the fact that we deliberately asked gynecologists from the same hospital and department to be assessors. We therefore suggest further research to explore the reliability of videotaped laparoscopy evaluations for endometriosis by a group of assessors from another medical center. However, because we found a poor level of agreement between the adhesions scorings, first of all studies are required to improve internal consistency of these evaluations.

In conclusion, for endometriosis, the use of videotaped laparoscopies seems to be justified because evaluations during (videotaped) laparoscopies proved to be reliable. Special attention must be given to the assessments of ovarian lesions because observers tend to disagree on the severity and extent of the endometriotic disease, which results in disagreements on the stage of the disease with therapeutic consequences. Regarding adhesions, the evaluations during videotaped laparoscopy were not reliable. These findings indicate that, in the case of adhesions, evaluations during videotaped laparoscopies should be interpreted with caution. Therefore, in court or when second opinions are requested regarding infertility or patients with chronic pelvic pain, one cannot rely on videotaped findings only. If advice on any therapeutic consequences is warranted, repeated surgery (i.e., diagnostic laparoscopy) may be necessary.

Acknowledgements
The authors thank Wouter Droog, MD, for his help in data collection and Anja Greeven, psychologist, for her contribution preparing the first draft of this manuscript.