Inter-rater reliability study of the modified Oxford Grading Scale and the Peritron manometer

Cristine Homsi Jorge Ferreira a,*, Patrícia Brentegani Barbosa a, Flaviane de Oliveira Souza a, Flávia Ignácio Antônio a, Maíra Menezes Franco a, Kari Bø b

a Faculty of Medicine of Ribeirão Preto, Department of Biomechanics, Medicine and Rehabilitation of Locomotor System, Course of Physiotherapy, University of São Paulo, Av. Bandeirantes, 3900, Monte Alegre, 14049-900, Ribeirão Preto, SP, Brazil
b Norwegian School of Sport Sciences, Department of Sports Medicine, Oslo, Norway

Abstract

Objective To evaluate the inter-rater reliability of the modified Oxford Grading Scale and the Peritron manometer.

Design All participants were evaluated twice, first by one examiner and 30 days later by a second examiner. Measurements of vaginal squeeze pressure were compared with the results from the palpation test.

Participants Nineteen women with a mean age of 23.7 years (range 21 to 28 years).

Results Inter-rater reliability for vaginal palpation was fair (κ = 0.33, 95% confidence interval 0.09 to 0.57). Using the Peritron manometer, the difference between examiners was less than 10 cmH 2O in 11 of the 19 (58%) cases. The palpation test did not differentiate between weak, moderate, good and strong muscle contractions. This study found fair inter-rater reliability for the modified Oxford Grading Scale and moderate inter-rater reliability for the Peritron manometer.

Conclusions The inter-rater reliability of vaginal squeeze pressure measurement using the Peritron manometer is acceptable and can be used in re-evaluations performed by different examiners in clinical practice. However, for research purposes, the ideal situation would be for a single examiner to assess and re-assess the subject. Vaginal palpation is important in the clinical assessment of correctness of a pelvic floor muscle contraction, but this study does not support the use of the modified Oxford Grading Scale as a reliable and valid method to measure and differentiate pelvic floor muscle strength.

© 2010 Chartered Society of Physiotherapy. Published by Elsevier Ltd. All rights reserved.

Keywords: Assessment; Inter-rater reliability; Palpation; Pelvic floor; Squeeze pressure; Strength

Introduction

There is Level 1 evidence that pelvic floor muscle training is effective for the treatment of mixed and stress urinary incontinence [1,2]. When conducting pelvic floor muscle training programmes, it is essential to measure their function and strength development [3]. Several vaginal palpation rating scales have been used in clinical practice [3], but the most commonly used tool in physical therapy seems to be the modified Oxford Grading Scale [4,5].

Whether a measurement tool should be used in clinical practice or in research depends on its reliability [6]. Inter-rater reliability refers to one researcher measuring the same procedure in the same subject twice, while inter-rater reliability refers to two or more clinicians or researchers conducting measurements of the same subject [6]. Some studies have reported high intra-rater reliability for vaginal digital assessment [7–9], but inter-rater reliability varies between studies [10–12]. Although the modified Oxford Grading Scale is considered by some to be an easy test and well tolerated by patients [8], others question its use for scientific purposes [9]. Recently, Slieker-ten Hove et al. [13] introduced a new digital assessment scale based on definitions of outcome measures of existing scales according to the standardised terminology of the International Continence Society, showing satisfactory face validity and intra-rater reliability but low inter-rater reliability.
Pelvic floor muscle strength can be measured with a manometer or a dynamometer. However, since dynamometers are not commercially available [3,14], strength is commonly measured by digital muscle testing and manometers [15–18]. The Peritron manometer (Cardio-Design, Victoria) has been found to have good intra-rater reliability [9]. However, few reports are available regarding its inter-rater reliability. Hundley et al. [19] reported that the Peritron manometer had high inter-rater reliability, but not all subjects were evaluated by the same examiners, and the authors stated that, in the purest sense, an assessment of inter-rater reliability requires that all subjects be examined by all examiners [19]. Thus, there seems to be a lack of studies truly evaluating the inter-rater reliability of this device.

If more than one examiner is to assess the participants in clinical studies, it is essential to assess the inter-rater reliability of the measurement method [3]. To date, there seems to be an immediate need for inter-rater reliability studies of both the modified Oxford Grading Scale and the Peritron manometer. In addition, Bo and Finckenhagen [12] questioned the ability of vaginal palpation to differentiate correctly between different degrees of pelvic floor muscle strength, and their results need further investigation.

The primary aim of this study was to evaluate the inter-rater reliability of the modified Oxford Grading Scale and the Peritron manometer. A secondary aim was to compare the scores obtained with the modified Oxford Grading Scale and the Peritron manometer for squeeze pressure measurement.

Materials and methods

Study design

This was a test–retest study assessing the inter-rater reliability of the modified Oxford Grading Scale and the Peritron manometer.

Participants

An initial sample of 20 female university students participated in the study. All participants signed an informed consent form, and the study was approved by the institutional research ethics committee.

Inclusion criteria were: age between 18 and 35 years, normal body mass index (<25 kg/m²), nulliparous and non-pregnant, and no gynaecological complaints or disease verified by at least one medical appointment in the past 12 months. Exclusion criteria were: pelvic organ prolapse or reconstructive pelvic surgery, symptoms of vaginal infection, intolerance to condoms, allergy to the gel used in the procedure, and involvement in pelvic floor muscle training. One participant reported that she had trained her pelvic floor muscles between the two evaluations, and was excluded from the study. Thus, the final sample consisted of 19 women.

Assessment tools

The modified Oxford Grading Scale [4] quantifies pelvic floor muscle strength as: 0, no contraction; 1, flicker; 2, weak; 3, moderate; 4, good; and 5, strong.

The Peritron manometer measures vaginal squeeze pressure through a conical sensor covered with a medical silicone rubber sheath (Fig. 1). The sensor is connected to a handheld microprocessor with a latex tube, allowing measurement of squeeze pressure in centimetres of water (cmH₂O). The occlusive pressure readings from a manometer are a surrogate measure of strength.

Interventions

An interviewer (PBB) questioned the participants about their age, weight, height, use of oral contraceptives, level of physical activity, participation in pelvic floor muscle training, parity and urinary incontinence.

After verifying that the participants were eligible to take part in the study, pelvic floor muscle strength was assessed. Examinations were conducted with the participants in a lithotomy position. All women were evaluated twice, first by one examiner and 30 days later by a second examiner. Both examiners (FDOS and MMF) were trained in this specific protocol of pelvic floor muscle assessment by a physiotherapist with 11 years of clinical experience and supervision of pelvic floor physiotherapists. Moreover, both examiners had equivalent
skills in pelvic floor physiotherapy and worked exclusively in women’s health. Examiner 1 (FDOS) had 4 years of clinical experience and Examiner 2 (MMF) had 3 years and 6 months of clinical experience. The interviewer (PBB) remained in the room to ensure that the same procedures were performed by the two examiners. The examiners were blinded to the results obtained by each other.

Vaginal palpation was performed using two fingers. The ability to contract and relax the pelvic floor muscles was first evaluated by palpation, asking the subject to pull her pelvic floor muscles in and up as strongly as possible and then to relax them completely. When a correct contraction was verified, the examiner scored it according to the modified Oxford Grading Scale [7].

After vaginal palpation, the Peritron manometer was placed with the middle of the probe 3.5 cm inside the vagina [17]. The device was calibrated to zero before each measurement. The vaginal sensor was not inflated. According to the manufacturer, inflation is an optional resource that can reduce the sensitivity of the sensor’s response. The women were instructed to undertake three maximal pelvic floor muscle contractions sustained for 5 seconds with an interval of 30 seconds, as reported in a previous study [20]. Only contractions with visible inward movement of the perineum were considered to be valid [21]. Co-contraction of the gluteal and hip adductor muscles was discouraged, as was the Valsalva manoeuvre [5,9,22]. At the end of the second evaluation, the participants were asked if they had done any pelvic floor muscle training between the two evaluations.

Statistical analysis

Background variables are reported as frequencies and percentages. The Kappa coefficient was used to assess the inter-rater reliability of the modified Oxford Grading Scale [23].

The mean peak value of the three contractions was used to analyse the inter-rater reliability of the Peritron manometer [20,24]. Bland and Altman limits of agreement [25–29] were used to compare the measurements obtained by the two examiners using the Peritron manometer, and the Pearson correlation test [30] was used to compare the values obtained with the Peritron manometer and the modified Oxford Grading Scale.

Results

The mean age of the participants was 24 years (range 21 to 28 years). All the participants were nulliparous White women who lived in Ribeirão Preto city, Brazil. Mean body mass index was 21.3 (range 17.1 to 24.7). Seventeen of the 19 participants (89%) were using an oral contraceptive, and six (31%) were undertaking regular general physical exercise. None of the subjects included in the analysis had done pelvic floor muscle training before participating in the research project or between the two evaluation points. One woman reported sporadic symptoms of urge urinary incontinence. During palpation, all women included in the study were able to voluntarily contract and completely relax their pelvic floor muscles. No women had hypertonic pelvic floor muscles.

Table 1 shows the pelvic floor muscle strength of each participant classified by the two examiners using the modified Oxford Grading Scale. The weighted Kappa was fair [$\kappa = 0.33$, 95% confidence interval (CI) 0.09 to 0.57]. There was agreement between the examiners for nine of the 19 subjects (47%). No participants were classified as Category 0 or 1 by either examiner. Examiner 2 scored three of the 19 (16%) participants as Category 5; Examiner 1 did not score any participants as Category 5. In 10/19 (53%) evaluations, Examiner 2 scored the participant as a higher category than Examiner 1.

The mean vaginal squeeze pressure of three contractions for the whole group was 44 cmH2O (95% CI 35.5 to 52.5) in the first evaluation and 46.7 cmH2O (95% CI 37.9 to 55.5) in the second evaluation. The difference between the evaluations was not statistically significant ($P = 0.65$, 95% CI $-14.46$ to 9.13).

Table 2 shows the mean of three maximal contractions for each participant in the first and second examinations using the Peritron manometer.

Table 3 shows the mean muscle strength (cmH2O) with 95% CI measured using the Peritron manometer for the six categories assessed by the modified Oxford Grading Scale for both examiners. Examiner 1 did not score any participants as Categories 0, 1 or 5, and Examiner 2 did not score any participants as Categories 0 or 1.
Table 2
Mean of three maximal pelvic floor muscle contractions (cmH2O) of individual participants by two different examiners using the Peritron manometer, n = 19.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Examiner 1</th>
<th>Examiner 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27.1</td>
<td>30.1</td>
</tr>
<tr>
<td>2</td>
<td>33.3</td>
<td>37.3</td>
</tr>
<tr>
<td>3</td>
<td>70.4</td>
<td>80.9</td>
</tr>
<tr>
<td>4</td>
<td>41.4</td>
<td>40.4</td>
</tr>
<tr>
<td>5</td>
<td>36.8</td>
<td>31.9</td>
</tr>
<tr>
<td>6</td>
<td>44.3</td>
<td>61.7</td>
</tr>
<tr>
<td>7</td>
<td>50.5</td>
<td>26.8</td>
</tr>
<tr>
<td>8</td>
<td>50.1</td>
<td>51.5</td>
</tr>
<tr>
<td>9</td>
<td>66.8</td>
<td>54.6</td>
</tr>
<tr>
<td>10</td>
<td>37.8</td>
<td>56.6</td>
</tr>
<tr>
<td>11</td>
<td>42.7</td>
<td>23.7</td>
</tr>
<tr>
<td>12</td>
<td>76.4</td>
<td>84.1</td>
</tr>
<tr>
<td>13</td>
<td>47.2</td>
<td>44.0</td>
</tr>
<tr>
<td>14</td>
<td>20.7</td>
<td>27.9</td>
</tr>
<tr>
<td>15</td>
<td>64.0</td>
<td>39.3</td>
</tr>
<tr>
<td>16</td>
<td>25.6</td>
<td>64.3</td>
</tr>
<tr>
<td>17</td>
<td>16.5</td>
<td>25.8</td>
</tr>
<tr>
<td>18</td>
<td>23.9</td>
<td>41.7</td>
</tr>
<tr>
<td>19</td>
<td>60.5</td>
<td>63.8</td>
</tr>
</tbody>
</table>

Mean for Examiner 1: 44 cmH2O (95% confidence interval 35.5 to 52.5).
Mean for Examiner 2: 46.7 cmH2O (95% confidence interval 37.9 to 55.5).

Fig. 2 shows the limits of agreement of pelvic floor muscle squeeze pressure measurements obtained with the Peritron manometer by the two examiners using the mean of three maximal readings. In 11 of 19 (58%) cases, the difference between the examiners was less than 10 cmH2O. In 7 of 19 (37%) cases, the difference was more than 15 cmH2O. It is estimated that in 95% of each test, the examiners will not exceed the limits of agreement, shown in the Fig. 2.

Fig. 3 shows a scatter plot with the values scored by the two examiners for the six categories of the modified Oxford Grading Scale, and correlation with the values for pelvic floor muscle squeeze pressure measurements obtained with the Peritron manometer. The results of Pearson’s correlation test were $r = 0.25$ (95% CI $-0.23$ to 0.63) for Examiner 1 and $r = 0.51$ (95% CI 0.08 to 0.78) for Examiner 2.

Table 3
Mean muscle strength (cmH2O) with 95% confidence intervals (CI), measured with the Peritron manometer, in the six categories assessed by the modified Oxford Grading Scale by the two examiners.

<table>
<thead>
<tr>
<th>Category</th>
<th>Examiner 1</th>
<th>Examiner 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>No contraction</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Flicker</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Weak</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>Moderate</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Good</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Strong</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Examiner 2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No contraction</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Flicker</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Weak</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Moderate</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Good</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Discussion
This study found fair inter-rater reliability for the modified Oxford Grading Scale, and moderate inter-rater reliability for the Peritron manometer. There was agreement between the examiners for nine of 19 subjects (47%) using the modified Oxford Grading Scale. Using the Peritron manometer, the difference between the examiners was less than 10 cmH2O in 11 of 19 (58%) cases. There were no differences between the Oxford Grading Scale categories compared with the results obtained with the Peritron manometer.

The results of studies evaluating inter-rater reliability of other palpation scoring systems and squeeze pressure using Spearman’s correlation rho range from $r = 0.60$ to $r = 0.90$ [13,19,31–34]. The findings of Bo and Fincken hagen [12] showed only acceptable (fair) inter-rater reliability for the modified Oxford Grading Scale using Cohen’s Kappa (0.37), despite a higher Spearman’s rho value (0.70). The present findings were similar using Cohen’s Kappa (0.33). Spearman’s rho was not used in the present study as it is not able to reveal systematic variations between two examiners, and may overestimate reliability, as shown by Bo and Fincken hagen [12].

The modified Oxford Grading Scale was used in this study because it is commonly used in clinical physical therapy. After this study was concluded, Slieker-ten Hove et al. [13] published a face validity and reliability study of the first digital assessment scheme of pelvic floor muscle function to conform with the standardised terminology of the International Continence Society. To evaluate voluntary pelvic floor muscle contraction and estimate strength, this scale used the
Fig. 3. Values estimated by the two examiners for the six categories of the modified Oxford Grading Scale, and correlation with the values obtained for pelvic floor muscle strength with the Peritron manometer (cmH₂O).

Brink score [32]. The authors stated that inter-rater reliability of this new scale was generally disappointing, but some items were good including the evaluation of voluntary contraction with the Brink score (κ = 0.64).

Although the modified Oxford Grading Scale allows the assessment of other aspects of pelvic floor muscle function, the only function that could be compared in this study was the strength (squeeze), since a perineometer was used to test the criterion validity of this scale [12]. The modified Oxford Grading Scale and manometry do not measure exactly the same aspects. According to Bo et al. [6], one of the difficulties of measurement using the modified Oxford Grading Scale is that it produces one value for occlusion and lift in one scale. The last three categories of the scale require the examiner’s palpating fingers to be sufficiently sensitive to notice not only occlusion but also the lift component. The lack of complete equivalence of these two parameters (squeeze and lift) measured by the two assessment tools may explain the lack of linearity between them and the low agreement obtained by the two examiners, especially for the last two categories of the modified Oxford Grading Scale, in the present study.

The results of muscle evaluation depend on the experience of the testers and the position of the subject being tested [12]. The authors controlled these factors as much as possible, and the two examiners were skilled physical therapists with similar experience in conducting this type of examination. In addition, they were blinded to each other’s results. All the participants were instructed in pelvic floor muscle contraction, and only correct contractions with perineal elevation were accepted [16,21]. Some studies have used the highest of three measurements for vaginal squeeze pressure to determine intra- or inter-rater reliability [2,9,12], while others, including the present study, have used the mean of three measurements due to possible variations related to learning and fatigue [20,24].

For inter-rater reliability of the Peritron manometer, this study found that the differences between examiners were less than 10 cmH₂O in 11 of 19 (58%) cases, indicating acceptable agreement. However, in seven of 19 (37%) cases, the differences were more than 15 cmH₂O. The only other study identified that assessed the inter-rater reliability of the Peritron manometer found high correlation between the results of the two examiners [19]. The study included 100 women (18 nulliparous and 82 parous). However, a subgroup analysis showed that correlation between examiners was unaffected by parity [19]. The authors explained issues that may have influenced their results, including the short interval between the two examinations and the fact that the examinations were performed by multiple examiners. Furthermore, not all subjects were evaluated by the same examiners [19]. The statistical methods used in the present study were different from the study mentioned above, making comparison difficult. Bland and Altman’s limits of agreement were used in the present study as it has been demonstrated that the use of correlation tests is inappropriate for the determination of reliability and may overestimate the results [25–27].

In the present study, all subjects were examined by the same two examiners, with an interval of 30 days between the two evaluations. This was done to prevent fatigue from influencing the measurements [9], but mostly for practical reasons. A shorter interval between the two examinations would be better, but it was impracticable for the participants to return to the place where examinations were performed before 30 days. This long interval may have given them time to train their pelvic floor muscles, although they were told not to do so between the two evaluations. Exercise between the two tests was an exclusion criterion, and one woman was excluded because she reported that she had trained. For the whole group, the mean values of vaginal squeeze pressure assessed by the two examiners did not differ, but Examiner 2 always scored the participants with higher categories on the modified Oxford Grading Scale than Examiner 1. Given this, it is not possible to guarantee that no learning effect occurred, although an increase in muscle strength is only expected after 5 months of intensive pelvic floor muscle training [35]. Since the results can be time- and rater-dependent, and given that
the time effect cannot be separated from the inter-rater effect, this long interval represents a limitation of the study.

The results of this study agree with the findings of Bo and Finckenhagen [12] showing that palpation scores using the modified Oxford Grading Scale did not differentiate pelvic floor muscle strength. The same methodology was used although the vaginal pressure manometers were different. The use of different types of perineometer generates different results that should not be compared [20,36]. In contrast, Isherwood and Rane [10] found good agreement between the modified Oxford Grading Scale and the PFX9100C perineometer (Cardio-Design, Victoria). However, the perineometer they used reported squeeze pressures on a 0 to 12 point scale, and in contrast with the present study, one examiner only used the modified Oxford Grading Scale and the other performed the evaluations with the perineometer. Other authors have found good correlation between different vaginal palpation scales and vaginal squeeze pressure [8,19,36,37]. A possible explanation of why the present study did not find any difference between the scores for vaginal palpation and vaginal squeeze pressure could be due to the small sample size and to the fact that some categories of the modified Oxford Grading Scale were not scored by the examiners. Another limitation of the present study is that the findings may only be valid for nulliparous women at a very young age, and the results may differ in older women with pelvic floor disorders.

Although the data analysis did not indicate that the group of women evaluated had generally weak pelvic floor muscles, more than 35% had their pelvic floor muscle contraction scored as weak or moderate by the two examiners. Dietz et al. [38] found that almost half of young and nulliparous women contracted their pelvic floor muscles unsatisfactorily or not at all unless they received instructions [38]. This indicates that the evaluation of pelvic floor muscles in young nulliparous women is essential, not only in research aiming to determine normal values of pelvic floor muscle strength, but also to evaluate preventive guidelines related to improvement of the ability to contract the pelvic floor muscles and to pelvic floor muscle training programmes in this population.

Despite the small sample size, the excessive homogeneity of the sample and the long interval between evaluations, the present results indicate that the inter-rater reliability of the Peritron manometer is acceptable and it can be used in re-evaluations performed by different examiners in clinical practice. However, for research purposes, the ideal situation is for the same examiner to assess and re-assess subjects. Despite the fact that the modified Oxford Grading Scale and the Peritron manometer do not evaluate exactly the same aspects of pelvic floor muscle function, the Oxford Grading Scale was not able to classify the degree of contraction correctly when compared with manometer readings. Vaginal palpation is important in assessing the correctness of a pelvic floor muscle contraction, but this study does not support the use of the modified Oxford Grading Scale as a reliable and valid method to measure pelvic floor muscle strength. Further test–retest studies in larger and more heterogeneous samples are warranted to substantiate these findings. The development of new pelvic floor muscle strength assessment tools and refinement of the existing tools is essential for reliable evaluations performed by more than one examiner.

Acknowledgements

The authors gratefully acknowledge the Foundation for the Support of Research of São Paulo State and Professor Edson Zangiacomi Martinez from the Center for Quantitative Methods for valuable advice with the statistical analysis.

Ethical approval: Institutional Research Ethics Committee of the University Hospital of Faculty of Medicine of Ribeirão Preto- University of São Paulo (FMRP-USP) (Protocol No. 12188/2007).

Conflict of interest: None declared.

References


