INTERTESTER RELIABILITY AND DIAGNOSTIC VALIDITY OF THE
CERVICAL FLEXION-ROTATION TEST

Toby M. Hall, PT, MSc,a Kim W. Robinson, PT, BSc,b Osamu Fujinawa, PT, PhD,c
Kiyokazu Akasaka, PT, PhD,d and Elizabeth A. Pyne, PT, MT)e

ABSTRACT

Objective: This article evaluates reliability and diagnostic validity of the cervical flexion-rotation test (FRT) to discriminate subjects with headache because of C1/2 dysfunction. In addition, this study evaluates agreement between experienced and inexperienced examiners.

Methods: These were 2 single blind comparative measurement study designs. In study 1, 2 experienced blinded examiners evaluated the FRT in 10 asymptomatic controls, 20 subjects with cervicogenic headache (CeH) where C1/2 was the primary dysfunctional level, and 10 subjects with CeH but without C1/2 as the primary dysfunctional level. In study 2, 2 inexperienced and 1 experienced blinded examiners evaluated the FRT in 12 subjects with CeH and 12 asymptomatic controls. Examiners were required to state whether the FRT was positive and also to determine range of rotation using a goniometer. An analysis of variance with planned orthogonal comparison, single measure intraclass correlation coefficient (2,1), and Bland-Altman plot were used to analyze FRT range of rotation between the examiners. Sensitivity, specificity, and examiner agreement for test interpretation were analyzed using cross tabulation and κ.

Results: In study 1, sensitivity and specificity of the FRT was 90% and 88% with 92% agreement for experienced examiners (P < .001). Overall diagnostic accuracy was 89% (P < .001) and κ = 0.85. In study 2, for inexperienced examiners, FRT mobility was significantly greater than for experienced examiners, but sensitivity, specificity, agreement, and κ values were all within clinically acceptable levels.

Conclusions: The FRT can be used accurately and reliably by inexperienced examiners and may be a useful aid in CeH evaluation. (J Manipulative Physiol Ther 2008;31:293-300)

Key Indexing Terms: Reproducibility of Results; Headache; Diagnosis; Physical Examination

Cervicogenic headache (CeH) has been classified by the International Headache Society (IHS)1 and is said to account for 15% to 20% of all chronic and recurrent headaches.2 Individuals report reduced quality of life3 and experience considerable restriction of daily function and emotional distress.4 There is encouraging evidence that CeH can be successfully managed by physical treatment.5,6 The IHS7 defined CeH as head pain, referred from a source in the neck that may arise from a variety of upper cervical spine structures.7,8 People who have CeH also complain of associated neck pain and restriction of neck movement.9 Unfortunately, neck pain is also a feature of other headache forms; hence, accurate diagnosis, by physical examination, is required to identify CeH.10 Although it has been suggested that CeH involves disorder of the articular, muscular, and neural tissue systems, Zito et al10 determined that the presence of upper cervical joint dysfunction most clearly identified patients with CeH. It is believed that one dysfunctional level may be the source of primary symptoms, with C1/2 being the most commonly reported.11-13

Cervical joint dysfunction can be measured by manual examination.14 Manual examination has high sensitivity and specificity to detect the presence or absence of cervical joint dysfunction in neck pain and headache patients.15-17 However, these tests involve a high degree of skill on the part of the examiner, and their reliability has been
questioned. It has been suggested that evidence of low reliability for these tests may be a reflection of poor research methods used by the studies to investigate these tests.

The cervical flexion-rotation test (FRT) is a simplified form of manual examination purported to identify C1/2 dysfunction; however, there are no studies that definitely support this conjecture. In this test procedure, the cervical spine is fully flexed, in an attempt to isolate movement to C1/2, which has a unique ability to rotate in flexion. Normal range of rotation in end range flexion has been shown to be 44° to each side. In contrast, subjects with C1/2 dysfunction have an average of 17° less rotation. Ogince et al reported the FRT was positive, if range was limited to 32° or less. They also demonstrated that highly trained manual therapists using the FRT have high sensitivity (91%) and specificity (90%) in identifying subjects with CepH from asymptomatic controls or subjects with migraine with aura. A limitation of that study was that the comparative groups had no cervical involvement and the CepH group highly defined. Hence, the reported reliability, sensitivity, and specificity may be artificially high. Further studies are required to investigate the FRT in more heterogeneous samples.

In the Mulligan and other manual therapy concepts, dysfunction on the FRT is used as an indicator to apply a specific treatment technique in patients with CepH. Many physiotherapists learn the FRT on postgraduate courses with limited supervision. It is not known whether, having learned the technique, physiotherapists are able to apply the FRT with the same degree of reliability, sensitivity, and specificity as more experienced examiners.

The purpose of this study was 2-fold. Firstly, to determine the reliability, agreement, and validity of the FRT when used by experienced examiners evaluating a heterogeneous sample of subjects with CepH. Secondly, to determine the influence of examiner experience on reliability of measurement, range of motion, and agreement of interpretation of the FRT. It was hypothesized that, when compared to experienced examiners, inexperienced examiners would be less reliable, record different range of motion, and be less sensitive and specific at identifying FRT dysfunction in patients with CepH.

METHODS

This was a single-blind, comparative measurement study design, undertaken as 2 separate studies, as a matter of convenience and to avoid exacerbation of subjects by repeated application of the FRT by multiple examiners. Study 1 was to investigate agreement between experienced examiners and to investigate the validity of the FRT as a test of C1/2 dysfunction. Study 2 was to investigate the influence of examiner experience on reliability, sensitivity, and specificity of the FRT. The study design is illustrated in a flow chart shown in Figure 1.

In study 1, 2 experienced examiners, blinded to each other, evaluated the FRT in a cohort consisting of 3 subgroups. Group A had CepH with C1/2 dysfunction, and group B also had CepH but with dysfunctional levels other than C1/2. Group C were all asymptomatic controls. In study 2, 2 inexperienced and one of the previous experienced examiners from study 1, again blinded to each other, evaluated the FRT in 2 groups. Group D were subjects with CepH, and group E were all asymptomatic control. This study was approved by Curtin University (Perth, Western Australia) human research ethics committee. The rights of individuals were respected at all times. In addition, subjects were able to withdraw from the study at anytime and gave written informed consent before the study commencement.

Subjects

Subjects were recruited as a sample of convenience from physiotherapy and medical clinics and physiotherapy course attendees. Subjects in study 1 did not take part in study 2. On entering the study, subjects were allocated to a CepH group or asymptomatic control group based on the following criteria. Asymptomatic controls had no significant history of neck pain or headache and were excluded if they had a headache more than once per month and neck pain that had required treatment in the last year. Inclusion criteria for CepH were based on guidelines of the Headache Classification Subcommittee of the IHS together with the Cervicogenic
Headache International Study Group. These criteria were unilateral or side dominant headache without side shift, headache associated with neck stiffness or pain, headache present continuously, or episodically for the past 3 months at least once per week and between the ages of 18 and 66 years. Average duration of headache was 4.9 years (SD, 3.4 years). Average headache severity, based on an index of combined headache intensity, duration, and frequency, was 45/100 (SD, 10.7). Exclusion criteria were headache not of cervical origin (IHS, 2004) and headache with autonomic involvement, dizziness, or visual disturbance.

In addition to the above criteria, in study 1, subjects with CeH were further classified into 2 groups by an independent trained manual therapist using manual examination (but without using the FRT), which sought to identify the primary dysfunctional level, a process used in similar studies of CeH evaluation. Manual examination consisted of passive physiologic intervertebral motion tests, as well as posterior-anterior (PA) pressures. The PA pressure was applied to each vertebra in turn and any symptom response identified. In this way, it was possible to identify the most painful dysfunctional cervical motion segment. Manual examination was used, as no credible alternative “reference standard” exists to identify the presence of C1/2 dysfunction. Thus, 2 groups were identified, those with and without primary dysfunction at C1/2.

The flow chart in Figure 1 details the group allocations and sample size. In study 1, all subjects in group A would be expected to show restriction on the FRT and group B and C with no restriction. Sample size calculations were based on Hall and Robinson and Ogince et al, who reported a positive cutoff value of 32° for the FRT. To detect a 10° difference in FRT mobility, with α at .05, power of 80%, and an SD of 8, this study required at least 10 subjects per group. This number was increased to 20 per group in study 1, in anticipation of the greater variability in FRT mobility in the comparative group, in contrast to the previous studies that used only control subjects with no neck problems.

Procedures

A separate, independent researcher screened for entry-level eligibility, based on the inclusion and exclusion criteria, by an initial interview. Before measurement, each subject with CeH was required to fill out a questionnaire, which has been shown to be reliable, to obtain an index of headache severity (0-100) based on headache frequency, pain disability, and duration. Larger scores indicate greater severity. This index was to provide demographic data.

In study 1, before the FRT, an independent manual therapist (20 years of postgraduate experience) assessed the subjects with CeH using manual examination to determine which cervical segment was the primary dysfunctional level. Consequently, 2 accredited Mulligan concept teachers, with more than 15 years experience, blinded to each other and to the subject group allocation, both separately performed the FRT procedure (Fig 2). With the subject relaxed supine and the cervical spine passively maximally flexed, the head was passively rotated left and right. Range was determined either by the subject reporting the onset of pain or firm resistance encountered by the therapist, whichever came first. Each assessor performed one repetition of the FRT in each direction on one occasion only, as previous research has indicated no improvement in reliability, for experienced examiners, with repeated testing. The examiner was required to state whether the FRT was positive or negative. A positive state was based on an eyeballed, estimated limitation of more than 10° from the anticipated normal range of 44° (SD, 8°). The examiner then measured FRT mobility (Fig 2) using a cervical range of motion (CROM) device. The CROM is a floating compass (Plastimo Airguide Inc [Compasses], Buffalo Groove, IL) attached to the apex of the head by Velcro straps. The order of testing was alternated between examiners to prevent order effects.

In study 2, 2 inexperienced examiners and one of the examiners from study 1 were evaluated for agreement. The inexperienced testers were volunteers participating in a Mulligan concept postgraduate course. Both had no previous experience of using the FRT but were qualified physiotherapists, working in clinical practice for approximately 5 years. The FRT was taught to both examiners, followed by 1 hour of supervised practice. The next day, all 3 examiners assessed all 24 subjects in alternating order, to prevent order effects. The 2 inexperienced examiners performed the FRT twice to determine intratherapist reliability. The experienced examiner performed the FRT once only, as a previous study has shown this examiner to have an intraclass correlation coefficient of 0.92 for repeated measures of the FRT. As in study 1, each examiner was required to report whether the FRT state was positive or not based on eyeballed
interpretation of range and resistance (on the first occasion only), before measuring range of motion using the CROM.

Data Analysis

Statistical analysis was carried out using SPSS V12.0. (SPSS Inc, Chicago, IL). α was set at .05 for each analysis.

Study 1

Reliability. A single measure intraclass correlation coefficient (ICC) (2,1) and confidence interval (CI) were generated for range recorded during the FRT. An analysis of variance and planned orthogonal comparison were used to analyze FRT range of rotation between the examiners. In addition, the differences in range between the corresponding examiners were plotted against the average of the 2 measurements (Bland-Altman plot) to look for trends and systematic bias. The limits of agreement, between the 2 assessors, were calculated from the mean difference ±2 SD. Intertester reliability of the interpretation of the FRT (positive or negative) was calculated from a cross tabulation using κ.

Validity. The sensitivity and specificity of the FRT were analyzed using cross tabulation and were determined with a receiver operating characteristic (ROC) chart. To calculate the sensitivity and specificity of experienced examiners, examiners assessment for group A was compared with group B and C combined. The dichotomous variables used to calculate sensitivity and specificity were the therapists’ identification of the presence or absence of dysfunction on the FRT.

Study 2

Reliability. Intraclass correlation coefficient (2,1) and CI were calculated for inexperienced examiners measures of FRT mobility. Analysis of variance and planned orthogonal comparisons were used to analyze FRT range of rotation between the examiners. Bland-Altman charts were plotted to look for trends and systematic bias. Intertester reliability of the interpretation of the FRT was calculated using κ.

Validity. Sensitivity and specificity were analyzed as per study 1. To calculate the sensitivity and specificity of inexperienced examiners, results from the experienced examiner assessment of group D and E were compared with the inexperienced examiners.

RESULTS

Part One

Reliability. In study 1, for experienced examiners, inter-examiner ICC values for the FRT were 0.93 (CI, 0.87-0.96), indicating excellent reliability. For asymptomatic subjects (group C), average range of unilateral rotation, to the left and right, was 44° (SD, 3.1°) and 43° (SD, 2.8°), respectively. For subjects with CeH without C1/2 being the primary dysfunctional level (group B) average rotation was 42° (SD, 2.9°) and 40° (SD, 3.6°) for left and right, respectively. In contrast, for subjects with CeH and C1/2 as the primary dysfunctional level (group A), the range toward the most restricted side was 30° (SD, 6.9°). The difference in range of rotation between groups was significant (F(2,37) = 19.44; P < .001). Furthermore, there was a small but significant difference in range (4°) between group B and group C (P < .001). A Bland-Altman plot comparing the 2 experienced examiners is presented in Figure 3. This chart shows that all the points are inside the ±2 SD line, and there is an even distribution either side of the mean, illustrating excellent agreement between experienced examiners. Cross tabulation revealed that 92% of the time both testers agreed on their interpretation of the FRT (P < .001). The κ value for examiner agreement on outcome of the FRT was 0.85, again indicating excellent agreement.

Validity. Sensitivity is the test’s ability to obtain a positive test when the target condition is really present. Specificity is the test’s ability to obtain a negative test when the condition is really absent. A positive likelihood ratio indicates the increase in odds, favoring the condition, given a positive test result, and a negative likelihood ratio indicates the change in odds, favoring the condition, given a negative
test result.27 The frequencies used to calculate sensitivity, specificity, and likelihood ratios are given in Table 1. Table 2 outlines the sensitivity, specificity, positive, and negative likelihood ratios for each examiner. Mean sensitivity was 90% and specificity was 88% for the 2 experienced examiners (P < .001).

An ROC chart shows the relationship between sensitivity and specificity. The area under the curve represents the ability of the test to discriminate between the diseased and nondiseased state. The ROC chart revealed, that presented with a randomly chosen pair of patients, using the FRT, the experienced clinician is able to make the correct diagnosis 89% of the time (P < .001). In addition, coordinates on the ROC curve indicated that the test value that provides the highest sensitivity and the lowest 1-specificity is 33° (cutoff score). That is, if the FRT value is at least 33°, the test result is positive.

Study 2

Reliability. In study 2, ICC values for the FRT comparing the experienced with 2 inexperienced examiners were 0.84 (CI, 0.67-0.93) for examiner 1 and 0.76 (CI, 0.53-0.89) for examiner 2. Intraexaminer ICC values for the 2 inexperienced examiners were 0.84 (CI, 0.66-0.92) and 0.89 (CI, 0.76-0.95). Average range of unilateral rotation for the experienced and 2 inexperienced examiners for the asymptomatic and CeH group is shown in Table 3.

The FRT range of motion, toward the side of restriction, was significantly different between the examiners (F (2,33) = 3.8; P < .05). The difference between the experienced and 2 inexperienced examiners was 5° and 9° for examiner one and two, respectively (Fig 4). Both inexperienced examiners recorded significantly larger range than the experienced examiner for both the asymptomatic and CeH groups (P < .05), but there was no difference between the 2 inexperienced examiners (P = .32).

Bland-Altman plots comparing the experienced examiner with the 2 inexperienced examiners are presented in Figures 5 and 6. These charts show that only 1/24 (4%) of the points are outside the ±2 SD lines. Both inexperienced examiners appear to measure “high” in comparison to the experienced examiner, with the second examiner worse than the first. In addition, the SD of the difference between second and experienced examiner was much greater than the first.

Cross tabulation revealed that agreement with the experienced examiner with the 2 inexperienced examiners occurred 88% and 83% of the time (P < 0.05). The κ values for these examiners were 0.75 and 0.67 (P < .001).

Validity. The frequencies used to calculate sensitivity, specificity, and likelihood ratios, for the 2 inexperienced

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Experienced examiners</th>
<th>Inexperienced examiners</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 (CI)</td>
<td>2 (CI)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>90% (77-96)</td>
<td>83% (65-90)</td>
</tr>
<tr>
<td>Specificity</td>
<td>90% (77-96)</td>
<td>92% (73-98)</td>
</tr>
<tr>
<td>Positive likelihood ratio</td>
<td>9 (3.3-24.2)</td>
<td>10 (2.4-53.9)</td>
</tr>
<tr>
<td>Negative likelihood ratio</td>
<td>0.11 (0-0.3)</td>
<td>0.18 (0.1-0.5)</td>
</tr>
</tbody>
</table>

Table 2. Combined results of study 1 and 2

<table>
<thead>
<tr>
<th>Examiner</th>
<th>Rotation left°</th>
<th>Rotation right°</th>
<th>Restricted side°</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymptomatic</td>
<td>CeH</td>
<td>Asymptomatic</td>
<td>CeH</td>
</tr>
<tr>
<td>Experienced</td>
<td>44 ± 3.4</td>
<td>45 ± 2.2</td>
<td>26 ± 8.4</td>
</tr>
<tr>
<td>Inexperienced 1</td>
<td>48 ± 3.9</td>
<td>48 ± 2.9</td>
<td>31 ± 8.3</td>
</tr>
<tr>
<td>Inexperienced 2</td>
<td>51 ± 2.9</td>
<td>50 ± 3.8</td>
<td>35 ± 5.8</td>
</tr>
</tbody>
</table>

Mean and SD of mobility recorded during the FRT in subjects with CeH and asymptomatic controls. Range = mean ± SD (range, SEM).

Table 3. Results of study 2

Fig 4. Results of study 2. Mean range and 95% CI for range of the FRT toward restricted side for experienced and 2 inexperienced examiners. 1 indicates experienced examiner; 2 and 3, inexperienced examiners.
examiners, are given in Table 4. Sensitivity, specificity, and likelihood ratio values for these examiners are shown in Table 2.

DISCUSSION

The main findings of this study were that diagnostic accuracy, sensitivity, and specificity of the FRT were very high for experienced examiners. Furthermore, although inexperienced examiners reported greater range for the FRT than experienced examiners, sensitivity, specificity, and agreement were still within clinically acceptable levels.

Study 1 found excellent reliability for FRT mobility measures and examiner agreement of FRT interpretation. For experienced examiners, ICC values for intertester reliability were 0.93, and for examiner agreement, \( \kappa \) was 0.85. It is recommended that for most clinical tests in physical examination, the \( \kappa \) value should range between 0.4 and 0.6.\(^{28} \) The present study found \( \kappa \) was 0.85 indicating excellent agreement for FRT interpretation. These findings are consistent with previously published reports of high reliability and agreement for the FRT, when measured by experienced examiners.\(^{12,22} \)

The FRT mobility was 43° to the right and 44° to the left, for asymptomatic subjects, highly consistent with other studies of normal values for this measurement.\(^{12,20} \) Furthermore, FRT mobility was reduced by 14° in the presence of CeH with C1/2 dysfunction when compared to asymptomatic subjects or subjects with CeH with no C1/2 dysfunction. These results are consistent with Ogince et al,\(^{22} \) who reported a limitation of 19° and 17°, respectively. Clinically, a difference in range of 14° or more should be readily detectable.

This is the first study to report FRT mobility across different subgroups of subjects with CeH. We found a mean 4° reduction of FRT mobility, in subjects with CeH, whose primary dysfunctional level was not C1/2 (group B), when compared to asymptomatic controls (group C). It has been suggested that CeH is not usually a disorder that involves only one cervical motion segment.\(^{10} \) Although subjects in group B were classified as having pain arising primarily from a cervical level other than C1/2, it is probable that C1/2 may have been involved to some lesser extent; hence, the observed small statistically significant, but clinically insignificant, reduction of FRT mobility. The much greater mobility limitation in group A, all subjects with the primary dysfunctional level at C1/2, helps to confirm the use of the FRT in manual examination, in particular, as a component of segmental diagnosis and in CeH evaluation. An alternative explanation is that the FRT is not an isolated test of C1/2 motion and that other cervical levels contribute to this movement. Hence, restriction of movement at any cervical level may contribute to the limitation of FRT mobility. Further studies are required to investigate the specificity of the FRT to the C1/2 level.

This study found diagnostic accuracy of the FRT, for experienced examiners, was 89%; sensitivity, 90%; and specificity, 88%. A positive cutoff value, for a positive test, was found to be 33°. These findings are consistent with Ogince et al,\(^{22} \) who reported sensitivity and specificity of 91% and 90%, respectively, with an overall diagnostic accuracy of 91% and a similar positive cutoff value. These results indicate very high accuracy and therapist agreement,\(^{25} \) in determining the presence of C1/2 dysfunction compared with a reference standard of manual examination. As stated, Ogince et al evaluated the FRT in 3 different groups with only one group having cervical spine involvement. One criticism of that study, therefore, is that the accuracy of their results may be inflated because
The present study confirms the findings of Ogince et al.22 CeH with pain arising from cervical levels other than C1/2, in subjects with no cervical spine involvement as well as those subjects with the primary dysfunctional level using manual examination. No other method of determining the dysfunctional level was available, and this form of manual examination has been used in similar studies.10,12,16,22 An additional weakness of the study was that the subjects were a sample of convenience from a variety of sources. Future studies should investigate the diagnostic accuracy of the FRT in subjects presenting concurrently to clinical practice. This would enable a more accurate determination of the validity and diagnostic accuracy of the FRT.

**Conclusion**

This study shows that the FRT has a high level of examiner agreement for experienced examiners. Inexperienced examiners show lower levels of agreement, but this is still at acceptable levels. The FRT is a useful clinical measure of cervical movement impairment and can aid in the differential diagnosis of CeH.

### Practical Applications

- The FRT is accurate and reliable when used by experienced and inexperienced examiners.
- This study provides preliminary evidence for the validity of the FRT as a measure of C1/2 dysfunction.

---

**Table 4. Results of study 2**

<table>
<thead>
<tr>
<th></th>
<th>Experienced examiner FRT</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Inexperienced examiner FRT</td>
<td>10/10*</td>
<td>1/2*</td>
</tr>
<tr>
<td>Negative</td>
<td>2/2*</td>
<td>11/10*</td>
</tr>
<tr>
<td>Total</td>
<td>12/12*</td>
<td>12/12*</td>
</tr>
</tbody>
</table>

The frequencies used to calculate sensitivity and specificity for 2 inexperienced examiners. Data are expressed as inexperienced examiner 1/inexperienced examiner 2*. The experienced examiner’s interpretation of group D (n = 12) and E (n = 12) were used as the true positive and true negative.
REFERENCES