Using normalized EMG to define the nociceptive flexion reflex (NFR) 
threshold: Further evaluation of standardized NFR scoring criteria

Christopher R. France\textsuperscript{a,*}, Jamie L. Rhudy\textsuperscript{b}, Sarah McGlone\textsuperscript{a}

\textsuperscript{a} Department of Psychology, Ohio University, Athens, OH 45701, USA
\textsuperscript{b} Department of Psychology, University of Tulsa, Tulsa, OK 74104, USA

\textbf{A R T I C L E   I N F O}

Article history:
Received 18 March 2009
Received in revised form 4 June 2009
Accepted 17 June 2009

Keywords:
Nociceptive flexion reflex
Threshold
Measurement
Receiver operating characteristics

\textbf{A B S T R A C T}

The nociceptive flexion reflex (NFR) has been used as a psychophysiological tool to study spinal nociceptive processes in numerous clinical and experimental studies. Despite widespread use of the NFR, few attempts have been made to empirically test and compare different scoring criteria to detect the presence/absence of the reflex. The present studies were conducted to address this issue. Study 1 (N = 56 healthy participants) examined the reliability of 15 different scoring criteria that were examined in a previous report. Study 2 (N = 73 healthy participants) extended this work by examining normalized scoring criteria based on biceps femoris activity unrelated to noxious stimulation (reference contraction, maximal contraction). In both studies, receiver operating characteristics (ROCs) analyses were used to evaluate and compare different scoring methods. The results indicate that a number of different criteria were acceptable for defining an NFR threshold based on the area under the ROC curve and its statistical significance; however, NFR Interval z score [(NFR Interval Mean – baseline mean)/baseline SD] emerged as the scoring criterion with the greatest accuracy and with cut-points that are reliable across samples. These findings support the application of a common NFR scoring criterion to enhance direct comparison of results across different research laboratories and study samples.

© 2009 International Association for the Study of Pain. Published by Elsevier B.V. All rights reserved.

\textbf{1. Introduction}

For more than a half-century the nociceptive flexion reflex (NFR) has been a psychophysiological tool to study pain and nociception in hundreds of experimental and clinical studies [27]. The NFR is typically assessed by monitoring biceps femoris muscle electromyogram (EMG) activity following presentation of brief electrocutaneous stimulation over the ipsilateral sural nerve; the intensity of stimulation required to elicit the NFR is used as an index of nociceptive threshold [28]. Prior research has demonstrated that the NFR threshold is often highly correlated with pain threshold [6,17,31,32], making this a particularly useful measure in clinical and experimental studies of nociception and pain modulation.

Although the NFR paradigm has the apparent advantage of providing an objective and standardized assessment of nociception, in practice researchers have used a variety of different criteria to define the presence or absence of a reflex, and this has led to difficulties in making comparisons across studies and research laboratories. Recently, we compared different criteria and arrived at a recommendation regarding two metrics that proved to be the most reliable and accurate at discriminating the presence of an NFR [24]. Specifically, our findings supported both a standardized peak criterion [i.e., NFR Interval Peak z score, which was defined as (NFR Interval Peak response – baseline mean)/baseline SD] and a standardized mean criterion [i.e., NFR interval z score, which was defined as (NFR Interval Mean – baseline mean)/baseline SD] as superior to other possible criteria in two separate samples of healthy young adults.

The present study was designed to further examine the issue of NFR threshold definition by employing a normalization procedure to address anthropomorphic differences between recording sites and individuals (e.g., sensor placement, thickness of subcutaneous tissue, muscle mass) that can adversely affect comparisons of EMG muscle responses [33]. Perhaps the most common method of normalization involves calculating EMG activation as a percentage of maximum voluntary contraction, or percentage MVC [19]. Alternatively, if a maximum voluntary contraction is not possible or difficult to obtain for various reasons (e.g., it may produce pain or enhance existing pain), it is also possible to normalize EMG as a percentage of a submaximal reference voluntary contraction, or percentage RVC. For an NFR study, percentage MVC and percentage RVC would be calculated by having participants produce a maximum flexion of the biceps femoris or maintain a standardized angle of leg flexion, respectively. To address the viability of percentage MVC and percentage RVC criteria for defining NFR threshold, we examined data from two separate samples of healthy young adults undergoing NFR assessment in our laboratory.
data from the first sample we reassessed the accuracy of the NFR threshold criteria that we supported in our previous report [24]. Using data from the second sample we compared these criteria against percentage MVC and percentage RVC criteria.

2. Study 1: replication of previously validated scoring criteria for the NFR

2.1. Methods

2.1.1. Participants
Participants were 56 undergraduate students (23 female and 33 male) at Ohio University who received credit towards fulfilling a research requirement in a basic psychology course for their participation. The sample had a mean age of 18.9 years (SD = 0.7) and the majority of the sample (87.5%) was Caucasian.

2.1.2. Laboratory testing procedure
Participants were instructed to refrain from caffeine, nicotine, alcohol, and strenuous exercise for at least 4 h and from analgesic medication for 24 h prior to testing. To begin the session, the electrocutaneous stimulation and EMG recording sites were prepared and the required electrodes were attached. All electrode sites were cleaned and abraded with Omni Prep electrode paste and an impedance of less than 10 kOhm was achieved before proceeding, verified using a UFI Checktrode (model MKII). To record NFR activity, a differential EMG electrode was secured over the biceps femoris muscle of the left leg, 10 cm superior to the popliteal fossa, and a reference (common ground) electrode attached over the lateral epicondyle of the femur. Using a DelSys, Bagnoli-2 differential amplifier, EMG activity was sampled between 20 and 450 Hz, amplified \( \times 10,000 \), and recorded and processed using a CED Micro1401 analog-to-digital converter and Spike2 software. To elicit the NFR, electrocutaneous stimulation generated by a Digitimer D57A constant-current stimulator was applied over the retromalleolar pathway of the sural nerve of the left leg using a Nicolet bar electrode (a bipolar stimulating electrode with a fixed 30 mm inter-electrode distance) that was applied with cathode proximal. Each electrocutaneous stimulation was applied according to a variable interval schedule of 6 s (range 4–8 s) to decrease the likelihood of stimulus predictability.

Participants were then seated in a Human Touch Perfect Chair (model PCP2) adjusted to maintain knee flexion at approximately 60 degrees from horizontal. Then, participants received a series of four electrocutaneous stimulations of increasing intensity (0, 2, 4, and 6 mA) to help them acclimatize to electrocutaneous stimulation. NFR threshold was then determined using procedures described elsewhere [15,16,20,24]. Recordings obtained during NFR threshold assessment were not used to evaluate the scoring criteria and therefore are not discussed further.

Upon completion of the NFR threshold assessment, electrocutaneous pain threshold and tolerance levels were measured. Specifically, sural nerve stimulation trials were delivered as a volley of five 1 ms rectangular pulses with a 3 ms interpulse interval (total duration = 17 ms). Participants rated the perceived intensity of each stimulation using a Verbal Rating Scale (VRS) with anchors of 1 (sensory threshold), 25 (uncomfortable), 50 (painful), 75 (very painful), and 100 (maximum tolerable). Stimulation intensity began at 0 mA and increased in 2 mA steps until a maximum stimulation intensity of 40 mA was reached or the participant reported that they had reached their tolerance threshold (a VRS rating of 100). EMG data recorded during this last series of increasingly intense stimulations were then used to identify NFR threshold as described below.

2.1.3. Data files
A physiological data file was generated for every electrocutaneous stimulation trial during pain threshold and tolerance assessment. Each file included rectified biceps femoris EMG and a monitor signal for electrical stimuli, sampled at 2000 Hz and recorded 400 ms prior to and 1600 ms after each stimulation. A total of 673 files were collected from all 56 participants.

2.1.4. Procedures used by expert raters to identify NFR threshold
A program was developed using LabVIEW (National Instruments, Austin, TX) to display and score each physiological data file (see Fig. 1 for example waveforms). The program provided a graphic display of the rectified biceps femoris waveform and the electrocutaneous stimulation pulses and two experts familiar with the NFR independently scored each waveform. Instructions were to examine each waveform and determine whether a reflex occurred. Consistent with our prior study [24], expert ratings were used as the "gold standard" against which subsequent automated indices were compared. Specifically, using this graphic display, the operator was asked to indicate whether an NFR was present or absent for each waveform according to the following NFR definition: "An NFR exists if (1) at least one sizable difference peak occurs in the 90–150 ms post-stimulation window relative to the pre-stimulation baseline interval (−65 to −5 ms), but not if (2) activity in the 90–150 ms post-stimulation window mimics the pre-stimulation baseline".

Each of the waveforms was displayed with the scale of the y-axis set to −0.5 to 30 μV; however, in the event a trial contained voltages >30 μV, the scaling could be adjusted to −0.5 to 70 μV. Although a range of intervals (from 80 to 180 ms post-stimulation) has been used in prior investigations to identify the RII reflex, a more narrow 90–150 ms window avoids contamination of the RII reflex by the low-threshold cutaneous flexor reflex (RIF) which can precede 90 ms and by startle reactions and voluntary movements that can begin as early as 150 ms post-stimulation [9,10]. Accordingly, separate boxes were used in the graphic display to identify the pre-stimulation baseline interval (−65 to −5 ms) as well as the NFR interval (90–150 ms). Of the 673 individual files examined, the raters agreed on 654 (97.2%) regarding the presence (338 files, 50.2%) and absence (316 files, 46.9%) of an NFR (Interrater reliability: Kappa = .943, \( p < .001 \)). For trials with disagreement, raters met and were able to reach agreement on all of the files before ROC analyses were conducted. Six waveforms (0.9%) were identified by at least one rater as having significant noise during the pre-stimulation baseline. For all subsequent analyses, these waveforms were excluded. Thus, analyses were based on a final set of 667 (99.1% of original) waveforms; 345 (51.7%) of those files were said to have an NFR present.

2.1.5. Computation of automated EMG criteria
Separate from the expert ratings described above, an automated scoring program was used to compute a variety of biceps femoris EMG criterion variables (see Table 1). These criterion variables included 7 measures of absolute levels of EMG activity in the 90–150 ms post-stimulus NFR interval (i.e., NFR Interval Peak, NFR Interval Mean, NFR Interval AUC, NFR RMS, Number of Samples above 10 μV, Number of Samples above 20 μV, Number of Samples above 50 μV). The NFR Interval Peak was defined as the highest point (in μV) within the EMG curve between 90 and 150 ms post-stimulation. NFR Interval Mean was computed as the average level of EMG activity (in μV) within the same 90–150 ms window. NFR Interval AUC, or area under the curve, was computed as the sum of EMG activity (in μV) between 90 and 150 ms post-stimulation. NFR RMS, or root mean square, was computed as the square root of the average squared EMG activity (in μV) between 90 and 150 ms post-stimulation. Additional criteria included increases in...
Fig. 1. Examples of EMG waveforms that were judged by both expert raters as having a nociceptive flexion reflex (NFR) present or absent. The bottom panel illustrates stimulation onset/offset.

Table 1
Criterion variable definitions.

<table>
<thead>
<tr>
<th>Study 1</th>
<th></th>
<th>Study 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NFR Interval Peak</td>
<td>Peak voltage in NFR interval</td>
<td>%RVC (mean)</td>
<td></td>
</tr>
<tr>
<td>Baseline-Adjusted NFR Interval Peak</td>
<td>(NFR Interval Peak – baseline mean)</td>
<td>%MVC (mean)</td>
<td></td>
</tr>
<tr>
<td>NFR Interval Peak z score(^a)</td>
<td>(NFR Interval Peak – baseline mean)/baseline SD</td>
<td>%RVC (peak)</td>
<td>((NFR Interval Peak – baseline mean)/maximum contraction mean) × 100</td>
</tr>
<tr>
<td>NFR Interval Mean</td>
<td>NFR Interval Mean Voltage</td>
<td>%MVC (peak)</td>
<td>((NFR Interval Peak – baseline mean)/maximum contraction mean) × 100</td>
</tr>
<tr>
<td>Baseline-Adjusted NFR Interval Mean</td>
<td>(NFR Interval Mean – baseline mean)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NFR RMS</td>
<td>Root mean square of NFR interval</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline-Adjusted NFR RMS</td>
<td>(NFR Interval root mean square – baseline interval root mean square)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NFR Interval z score(^a)</td>
<td>(NFR Interval Mean – baseline mean)/baseline SD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NFR Interval AUC</td>
<td>NFR Interval total area under the curve</td>
<td>NFR Interval Cohen’s d(^d)</td>
<td>Statistical moment used to characterize the shape of NFR interval waveform</td>
</tr>
<tr>
<td>Baseline-Adjusted NFR Interval AUC</td>
<td>(NFR Interval total area under the curve – baseline total area under the curve)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NFR Interval Kurtosis</td>
<td></td>
<td>Number of Samples above 10 (\mu V)</td>
<td></td>
</tr>
<tr>
<td>Number of Samples above 10 (\mu V)</td>
<td>Number of NFR interval samples &gt; 10 (\mu V)</td>
<td>Number of Samples above 20 (\mu V)</td>
<td></td>
</tr>
<tr>
<td>Number of Samples above 20 (\mu V)</td>
<td></td>
<td>Number of NFR interval samples &gt; 20 (\mu V)</td>
<td></td>
</tr>
<tr>
<td>Number of Samples above 50 (\mu V)</td>
<td></td>
<td>Number of NFR interval samples &gt; 50 (\mu V)</td>
<td></td>
</tr>
</tbody>
</table>

Note: NFR, nociceptive flexion reflex; AUC, area under the curve; percentage RVC, percent of reference voluntary contraction; percentage MVC, percent of maximum voluntary contraction.

\(^a\) Refers to standardized criteria. \(z\) Scores are standardized because the standard deviation of baseline activity is used in the denominator, thus placing the variable in standard deviation units. The \(d\) score is standardized because the standard deviation of baseline activity and NFR activity is used in the denominator.
EMG activity in the NFR interval relative to the pre-stimulation baseline interval (Baseline-Adjusted NFR Interval Peak, NFR Interval Peak z score, Baseline-Adjusted NFR Interval Mean, NFR Interval z score, Baseline-Adjusted NFR Interval AUC, Baseline-Adjusted NFR RMS, NFR Interval Cohen’s d), and shape of the response waveform during the NFR interval (NFR Interval Kurtosis). The NFR Interval Kurtosis criterion was calculated to quantify the shape of the waveform during the 90–150 ms post-stimulus interval. Kurtosis is a statistical moment used to characterize the shape of a frequency distribution, with higher values representing more peaked or leptokurtotic distributions and lower values representing flatter or platykurtotic distributions. Interested readers are referred to our previous paper for details on how kurtosis was calculated [24].

2.1.6. Receiver operating characteristics (ROCs) analyses

To determine the criterion that best identified the presence of an NFR in Study 1, ROC analysis was conducted using the ROC procedure in SPSS 14.02. ROC analysis provides a means of comparing different variables that are intended for classification of a binary outcome (e.g., NFR present vs. NFR absent). To do so, the analysis must know a priori whether EMG trials are classified as having an NFR present or not based on a “gold standard”. The gold standard in the present study was the expert raters’ designation of each EMG trial as having an NFR present or not. The ROC analysis then compares different automatically generated EMG criteria (e.g., Baseline-Adjusted NFR Interval Peak, NFR Interval z score, Number of Samples above 50 μV) against the gold standard to determine how well each automated criterion is able to classify EMG trials. A ROC curve is generated for each automated criterion with 1-Specificity on the x-axis and Sensitivity on the y-axis. The area under the ROC curve quantifies how well each automated criterion performs. Specifically, when an automated criterion variable is unable to distinguish between the presence and absence of an NFR in the EMG trials, the area under the ROC curve equals 0.50. By contrast, an automated criterion that is more accurate at classifying the trials as having an NFR or not will have areas under the ROC curve closer to 1.0 (see Fig. 2). Furthermore, a significance test based on the asymptotic distribution of the test statistic is generated for the area under the curve to determine whether the automated criterion is able to classify the trials above chance levels. For the present study, an automated criterion was judged to be a “good” if the area under the ROC curve was high (closer to 1.0) and statistically significant at p < 0.05. Automated criteria that classify better than chance (i.e., are statistically significant) can be compared using the area under the ROC curve, but also the associated 95% confidence intervals for the area under the curve. A criterion was judged to be better than another if the area under the ROC curve was larger (greater accuracy) and the 95% confidence intervals for the areas were non-overlapping. All 667 trials were used for the ROC analyses.

ROC curves allow one to compare the classification accuracy of different automated criteria; however, it is also important to determine the cut-point for criteria that are judged to be accurate (i.e., what criterion value indicates the presence of an NFR). To determine cut-points, coordinate points for the ROC curves were generated. Coordinate points provided the specificity and sensitivity scores given a certain value of a criterion. For example, coordinate points might indicate that an NFR Interval z score of 1.38 is associated with a sensitivity of .955 and specificity of .943, whereas an NFR Interval z score of 1.00 is associated with a specificity of .979 and sensitivity of .902. Thus, coordinate points provide a way of comparing different cut-point in terms of the sensitivity and specificity associated with classifying EMG trials as having an NFR or not. For the present study, the optimal cut-point was determined to be the criterion value for which sensitivity = specificity.

Table 2
Receiver operating characteristics (ROCs) curve analyses on data from Study 1.

<table>
<thead>
<tr>
<th>Criterion variable</th>
<th>Area under the ROC curve</th>
<th>95% CI</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-Adjusted NFR Interval Peak</td>
<td>1.00</td>
<td>(0.99–1.00)</td>
<td>a</td>
</tr>
<tr>
<td>NFR Interval Peak</td>
<td>1.00</td>
<td>(0.99–1.00)</td>
<td>a</td>
</tr>
<tr>
<td>NFR Interval Peak z score</td>
<td>1.00</td>
<td>(0.99–1.00)</td>
<td>a</td>
</tr>
<tr>
<td>Baseline-Adjusted NFR RMS</td>
<td>1.00</td>
<td>(0.99–1.00)</td>
<td>a</td>
</tr>
<tr>
<td>NFR RMS</td>
<td>1.00</td>
<td>(0.99–1.00)</td>
<td>a</td>
</tr>
<tr>
<td>Baseline-Adjusted NFR Interval Mean</td>
<td>0.99</td>
<td>(0.99–1.00)</td>
<td>a</td>
</tr>
<tr>
<td>Baseline-Adjusted NFR Interval AUC</td>
<td>0.99</td>
<td>(0.99–1.00)</td>
<td>a</td>
</tr>
<tr>
<td>NFR Interval z score</td>
<td>0.99</td>
<td>(0.99–1.00)</td>
<td>a</td>
</tr>
<tr>
<td>NFR Interval Mean</td>
<td>0.99</td>
<td>(0.99–1.00)</td>
<td>a</td>
</tr>
<tr>
<td>NFR Interval AUC</td>
<td>0.99</td>
<td>(0.99–1.00)</td>
<td>a</td>
</tr>
<tr>
<td>NFR Interval Cohen’s d</td>
<td>0.97</td>
<td>(0.96–0.98)</td>
<td>b</td>
</tr>
<tr>
<td>Number of Samples above 10 μV</td>
<td>0.97</td>
<td>(0.95–0.98)</td>
<td>b</td>
</tr>
<tr>
<td>NFR Interval Kurtosis</td>
<td>0.86</td>
<td>(0.83–0.89)</td>
<td>c</td>
</tr>
<tr>
<td>Number of Samples above 20 μV</td>
<td>0.86</td>
<td>(0.83–0.88)</td>
<td>c</td>
</tr>
<tr>
<td>Number of Samples above 50 μV</td>
<td>0.68</td>
<td>(0.64–0.72)</td>
<td>d</td>
</tr>
</tbody>
</table>

Note: NFR, nociceptive flexion reflex; AUC, area under the curve; RMS, root mean square; CI, confidence interval. All area under the ROC curve estimates are significant (p < .001). Criteria are sorted from best to worst. The “Diff” column indicates which criteria are significantly different from one another – criteria sharing a letter are not significantly different (95% CIs are overlapping).
2.2. Results

2.2.1. ROC analyses

Fig. 2 depicts the ROC curves for selected automated criteria. As can be seen in Table 2, all scoring criteria performed well and all areas under the curve were significant at \( p < .05 \). Criteria are sorted from best to worst in terms of area under the ROC curve (although with rounding there are several ties). Peak variables (Baseline-Adjusted NFR Interval Peak, NFR Interval Peak, and NFR Interval Peak \( z \)-score) and RMS variables (Baseline-Adjusted NFR RMS, NFR RMS) performed best, with at least 99.5% of the area under the ROC curves. However, five other variables had at least 99% area under the ROC curve (Baseline-Adjusted NFR Interval Mean, Baseline-Adjusted NFR Interval AUC, NFR Interval \( z \)-score, NFR Interval Mean, NFR Interval AUC). Next best were NFR Interval Cohen’s \( d \) and Number of Samples above 10 \( \mu \)V with 97% area under the ROC curve. NFR Interval Kurtosis and Number of Samples above 20 \( \mu \)V were in the next lower tier of criteria, with 88% area under the ROC curve. Number of Samples above 50 \( \mu \)V had 68% area under the ROC curve.

2.2.2. Cut-point determination

Examples of graphs used to determine cut-points are depicted in Fig. 3. Coordinate points from the ROC curves for a subset of the above criteria were examined to determine the cut-point that optimized overall accuracy in defining a reflex (balanced sensitivity and specificity levels) for that criterion across all trials. Only those with greater than 90% area under the ROC curve were considered.

Furthermore, AUC variables and number of samples above 10, 20, and 50 \( \mu \)V were not reported because they are not conducive to generating cut-points that can be used across laboratories as their level is likely to vary as a function of EMG electrode design, quality of electrode application and surface preparation, participant muscle mass and subcutaneous adipose tissue, and sample rate. Table 3 presents the cut-points and the associated sensitivity and specificity values.

3. Study 2: evaluating NFR scoring criteria based on percentage MVC and percentage RVC

3.1. Methods

3.1.1. Participants

Participants were 73 undergraduate students at Ohio University who received credit towards fulfilling a research requirement in a basic psychology course for their participation. Forty-two were female and 31 were male. The mean age of participants was 19.00 ± 1.03 years. The majority (93.2%) of the participants were Caucasian.

3.1.2. Laboratory testing procedure

Using the same equipment and testing procedures described in Study 1, NFR threshold as well as pain threshold and tolerance was assessed. As in Study 1, EMG data recorded during pain threshold and tolerance assessment were used to identify NFR threshold as described below. A total of 1005 files were collected from all 73 participants.

---

Fig. 3. Plots of sensitivity (lighter line) and specificity (darker line) values (y-axis) by automated criteria values (x-axis) for four different NFR scoring criteria in Study 1. These graphs were used to determine the cut-points that balanced sensitivity and specificity levels (where the lines crossed), but can also be used by researchers to adjust the cut-points for these criteria to improve sensitivity or specificity.
Table 3
Cut-points for criteria that optimize overall accuracy in Study 1 and Study 2.

<table>
<thead>
<tr>
<th>Criterion variable</th>
<th>Study 1</th>
<th>Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SE and SP</td>
<td>SE and SP</td>
</tr>
<tr>
<td>Baseline-Adjusted NFR Peak</td>
<td>7.62 .99</td>
<td>7.50 .98</td>
</tr>
<tr>
<td>NFR Interval Peak z score</td>
<td>11.99 .98</td>
<td>12.32 .97</td>
</tr>
<tr>
<td>Baseline-Adjusted NFR RMS</td>
<td>1.47 .96</td>
<td>1.37 .96</td>
</tr>
<tr>
<td>Baseline-Adjusted NFR Interval Mean</td>
<td>0.95 .95</td>
<td>0.85 .94</td>
</tr>
<tr>
<td>NFR Interval z score</td>
<td>1.47 .95</td>
<td>1.42 .94</td>
</tr>
<tr>
<td>NFR Interval Cohen’s d</td>
<td>0.71 .91</td>
<td>0.63 .91</td>
</tr>
<tr>
<td>%MVC (peak)</td>
<td>– –</td>
<td>6.5 .95</td>
</tr>
<tr>
<td>%RVC (peak)</td>
<td>– –</td>
<td>53.0 .92</td>
</tr>
<tr>
<td>%MVC (mean)</td>
<td>– –</td>
<td>1.6 .89</td>
</tr>
<tr>
<td>%RVC (mean)</td>
<td>– –</td>
<td>14.7 .85</td>
</tr>
</tbody>
</table>

Note: percentage RVC, percent of reference voluntary contraction; percentage MVC, percent of maximum voluntary contraction; NFR, nociceptive flexion reflex; SE, sensitivity; SP, specificity; RMS, root mean square. The cut-point is the value of the criterion that maximizes both sensitivity and specificity. The sensitivity/specificity value listed in the SE and SP column corresponds to the cut-point. These values are the same, because cut-points were chosen based on criterion values for which sensitivity = specificity.

3.1.3. Procedures used to define NFR threshold
The procedures and raters used to define NFR threshold were the same as Study 1. Of the 1005 files, the two raters agreed on 969 (96.7%) regarding the presence (472 files, 47%) and absence (497 files, 49.5%) of an NFR (inter-rater reliability: Kappa = .92, p < .001). As in Study 1, raters met and were able to reach agreement on 100% of the files before ROC analyses were conducted. A total of 5 waveforms (0.5%) were identified by at least one rater as having significant noise during the pre-stimulation baseline. For all subsequent analyses, these waveforms were excluded. Thus, subsequent analyses were based on a final set of 1000 (99.5% of original) waveforms, with 476 (47.6%) of those files said to have an NFR present.

3.1.3.1. Additional scoring criteria. To determine percentage MVC and percentage RVC, participants were instructed to lay prone on a padded table prior to NFR assessment. The table had a padded, wooden bar fixed 28 cm above the table surface on one end. Participants were positioned so that, when bending the knee and resting their ankle above and then a second 5 s of EMG readings while holding the leg just above the bar. These two readings were averaged to get the “reference contraction mean” used to calculate percentage RVC in Table 1.

3.2. Results
3.2.1. ROC analysis
Data from Study 2 were analyzed to determine whether criterion variables from Study 1 performed consistently in a new independent sample and also to compare a new set of criteria based on pre-experiment biceps femoris EMG values (percentage RVC, percentage MVC). Like Study 1, criteria were compared using area

Table 4
Receiver operating characteristics (ROCs) curve analysis in Study 2.

<table>
<thead>
<tr>
<th>Criterion variable</th>
<th>Area under the ROC curve</th>
<th>95% CI</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>NFR Interval Peak</td>
<td>0.99</td>
<td>(0.99–1.00)</td>
<td>a</td>
</tr>
<tr>
<td>Baseline-Adjusted NFR Interval Peak</td>
<td>0.99</td>
<td>(0.99–1.00)</td>
<td>a</td>
</tr>
<tr>
<td>NFR RMS</td>
<td>0.99</td>
<td>(0.99–1.00)</td>
<td>a</td>
</tr>
<tr>
<td>NFR Interval Peak z score</td>
<td>0.99</td>
<td>(0.99–1.00)</td>
<td>a</td>
</tr>
<tr>
<td>Baseline-Adjusted NFR RMS</td>
<td>0.99</td>
<td>(0.99–1.00)</td>
<td>a</td>
</tr>
<tr>
<td>%MVC (peak)</td>
<td>0.99</td>
<td>(0.99–0.99)</td>
<td>a</td>
</tr>
<tr>
<td>NFR Interval AUC</td>
<td>0.99</td>
<td>(0.99–0.99)</td>
<td>a</td>
</tr>
<tr>
<td>NFR Interval Mean</td>
<td>0.99</td>
<td>(0.99–0.99)</td>
<td>a</td>
</tr>
<tr>
<td>Baseline-Adjusted NFR Interval AUC</td>
<td>0.99</td>
<td>(0.99–0.99)</td>
<td>a</td>
</tr>
<tr>
<td>Baseline-Adjusted NFR Interval Mean</td>
<td>0.99</td>
<td>(0.99–0.99)</td>
<td>a</td>
</tr>
<tr>
<td>NFR Interval z score</td>
<td>0.99</td>
<td>(0.98–0.99)</td>
<td>a,b</td>
</tr>
<tr>
<td>%RVC (peak)</td>
<td>0.98</td>
<td>(0.97–0.98)</td>
<td>b,c</td>
</tr>
<tr>
<td>NFR Interval Cohen’s d</td>
<td>0.97</td>
<td>(0.97–0.98)</td>
<td>b,c</td>
</tr>
<tr>
<td>%MVC (mean)</td>
<td>0.97</td>
<td>(0.96–0.98)</td>
<td>b,c</td>
</tr>
<tr>
<td>Number of Samples above 10 µV</td>
<td>0.95</td>
<td>(0.94–0.97)</td>
<td>c,d</td>
</tr>
<tr>
<td>%RVC (mean)</td>
<td>0.93</td>
<td>(0.92–0.95)</td>
<td>d</td>
</tr>
<tr>
<td>Number of Samples above 20 µV</td>
<td>0.83</td>
<td>(0.80–0.86)</td>
<td>e</td>
</tr>
<tr>
<td>NFR Interval Kurtosis</td>
<td>0.81</td>
<td>(0.78–0.84)</td>
<td>f</td>
</tr>
<tr>
<td>Number of Samples above 50 µV</td>
<td>0.65</td>
<td>(0.62–0.68)</td>
<td>g</td>
</tr>
</tbody>
</table>

Note: NFR, nociceptive flexion reflex; AUC, area under the curve; percentage RVC, percent of reference voluntary contraction; percentage MVC, percent of maximum voluntary contraction; CI, confidence interval. All area under the ROC curve estimates are significant (p < .001) except for Number of Samples above 50 µV (p = .44). The “Diff” column indicates which criteria are significantly different from one another – criteria sharing a letter are not significantly different (95% CIs are overlapping).
under the ROC curve and the associated 95% confidence intervals, with standardized criteria judged better than non-standardized criteria in the event of ties. As can be seen in Table 4, results were similar to those in Study 1 and all areas under the ROC curves were significant (p < 0.001). Fourteen criteria performed very well with greater than 96.5% area under the ROC curve. Included in these were the peak variables (NFR Interval Peak, Baseline-Adjusted NFR Interval Peak, NFR Interval Peak z score), the RMS variables (NFR RMS, Baseline-Adjusted NFR RMS), the AUC variables (NFR Interval AUC, Baseline-Adjusted NFR Interval AUC), and the mean NFR interval variables (NFR Interval Mean, Baseline-Adjusted NFR Interval Mean, NFR Interval z score, NFR Interval Cohen’s d). However, three of the new variables also performed well (percentage MVC-Peak, percentage RVC-Peak, percentage MVC-Mean). Two additional criteria had areas under the ROC curves greater than 90% (Number of Samples above 10 µV, percentage RVC-Mean), two had greater than 80% (Number of Samples above 20 µV, NFR Interval Kurtosis), and Number of Samples above 50 µV was 65%.

3.2.2. Evaluating cut-points

As in Study 1, the coordinate points from the ROC curve analyses were examined to determine the cut-points that optimized overall accuracy in defining a reflex. In addition to those variables examined in Study 1, the new scoring criteria (percentage RVC, percentage MVC) were also examined. Table 3 shows that most of the cut-points for Study 2 were similar to those derived in Study 1 (i.e., ±0.15), except for NFR Interval Peak z score which varied by only 0.33 U. Moreover, the new criteria suggested that peak biceps femoris EMG activity at NFR Threshold occurred at approximately 53% of reference voluntary contraction (percentage RVC) and 6.5% of maximum voluntary contraction (percentage MVC), whereas mean biceps femoris EMG activity at NFR Threshold occurred at 14.7% RVC and 1.6% MVC.

4. Discussion

The findings from the present study provide further support for the notion that mathematical criteria can be used to provide accurate and stable estimates of nociceptive flexion reflex threshold. More specifically, the results of the ROC analyses from both Studies 1 and 2 confirm our prior findings [24] that NFR Interval Peak z score was among the best criteria in terms of overall accuracy (largest area under the ROC curve). Further, among the criteria with the highest overall accuracy, the NFR Interval Peak z score is more likely to generalize across different participants and experimental settings due to the fact that this is a standardized criterion. It is important to note that 12/15 criteria tested in Study 1 and 14/19 criteria tested in Study 2 were statistically significant and exceeded 96.5% of the area under the ROC curve, suggesting many of these scoring methods are adequate for detecting a reflex. However, an important measure of the usefulness of a particular scoring method is the extent to which the ROC coordinate point that optimizes overall accuracy in defining a reflex (i.e., the cut-point that balances sensitivity and specificity) demonstrates stability across samples. As can be seen in Table 3, five of the six criteria that we examined had cut-points that varied by no more than 0.12 U between the first and second study samples. Further, 3 of the 4 criteria (Baseline-Adjusted NFR Interval Peak, Baseline-Adjusted NFR Interval Mean, and NFR Interval z score) had cut-points that varied less than 0.15 U from those observed in our prior report [24]. A notable exception was NFR Interval Peak z score, which varied 0.33 U between the current study samples and by up to 2.0 U from our prior report, suggesting that the cut-point for this scoring criterion may be less stable from sample to sample. Indeed, when the results of the present studies are combined with our previous report, the cut-point for NFR Interval z score had a standard deviation of .04, Baseline-Adjusted NFR Interval Peak had a standard deviation of .19, and Baseline-Adjusted NFR Interval Mean had a standard deviation of .08. However, NFR Interval Peak z score had a standard deviation of .92. For this reason, NFR Interval z score appears to be the best scoring criteria for both detecting a reflex and maintaining cut-point stability.

To our knowledge, Study 2 was the first attempt to define the presence/absence of an NFR by accounting for individual differences in biceps femoris contraction activity during unrelated stimulation. Prior to NFR testing, biceps femoris activity was recorded in response to a reference contraction (holding weight of leg up) and maximum contraction (pressing up against a fixed bar), and biceps femoris activity in response to noxious stimulation (in the 90–150 ms post-stimulation interval) was calculated as a percentage of these pre-experiment values (percentage RVC and percentage MVC, respectively). Results of the ROC analyses revealed that these measures can be used to accurately identify NFR threshold, with a percentage MVC–peak criterion providing the best overall performance followed by percentage RVC–peak (Table 4). Cut-point analyses suggested that NFR threshold tended to occur when peak responses were at least 6.5 of percentage MVC–peak or 53 of percentage RVC–peak.

Although replication of the present findings would be recommended before applying these criteria and specific cut-points, from a practical perspective this does not appear to be warranted. The normalization procedure requires additional time and effort on the part of the experimenter and the participant, and the resulting criterion measures do not yield greater accuracy in the prediction of NFR threshold than other criteria. Therefore the additional effort in collecting these measures may not be justified under most circumstances. Indeed, any of the Baseline-Adjusted criteria takes into account change from individual differences in “resting” biceps femoris activity, although not in terms of percentage change. Given this, under normal circumstances we cannot recommend these criteria over NFR Interval z score. It is possible, however, that there may be special circumstances under which these measures may be more desirable (e.g., clinical populations with large differences in participant anthropomorphics or laboratory settings with many different operators), and therefore additional investigation of these metrics may be advisable. One important issue to address in future investigations is whether the cut-points used to define NFR threshold (i.e., 6.5% and 53%) have better stability than other scoring criteria.

4.1. Implications and recommendations

Over the last few decades the NFR has emerged as an important biomarker of spinal nociceptive processes [27]. Indeed, it stands out as one of the only non-invasive biomarkers of spinal nociceptive processes and, along with pain-evoked event-related potentials (ERPs) [4,5,9] and neuroimaging (fMRI, PET) [7,21,22,30,34], is one of the few biomarkers of central nociceptive processes. However, NFR has a few advantages over these other methods. For example, research has shown that the stimulus intensity that reliably elicits the NFR (reflex threshold) correlates with subjective pain threshold [6,18,31,32] and the magnitude of the NFR correlates with pain intensity [6,26,32]. For these reasons, the NFR can be used in two ways: (1) as a measure of nociceptive threshold (i.e., 6.5% and 53%) have better stability than other scoring criteria.

Moreover, the signal-to-noise ratio is much higher for the EMG signals in NFR assessment than ERPs or neuroimaging. Thus, there can be less stress for the participant during measurement because fewer painful trials are needed. And, compared to neuroimaging, and to a lesser degree ERPs, NFR assessment is much less expensive.
because only a single bipolar AC amplifier is needed to record bi-
ceps femoris EMG activity. Thus, the NFR has much to offer to pain
researchers.

Nonetheless, to further the use of the NFR as a biomarker of no-
ceptive processes, it is important to standardize the methodology in
order to improve reliability and validity without compromising
practicality. Despite attempts by several different research groups
to generate replicable scoring criteria [1–3,8,11–15,26–29], there
has been little systematic research to compare the performance of
different criteria. The results of the present study, together with the
results of our previous report [24], indicate that NFR Interval z
score provides the best combination of overall accuracy and stability
as a standardized criterion of NFR threshold. Further, based on the
findings of the present report as well as our previous study, we rec-

ommend that researchers consider applying an NFR Interval z score
criterion of 1.4 SD in defining an NFR reflex. Adoption of a common
criterion for use with different populations, different equipment, and
in different settings would significantly enhance our ability to make
direct comparisons between studies in the future. It should be noted, however, that our efforts to identify a standardized NFR
criterion have been based on data collected from a relatively homo-
geneous population of young, healthy adults. Given that the NFR par-
cadigm is often used in research with older clinical populations, it is
advisable that further efforts be devoted to confirming our findings in
more heterogeneous samples.

Finally, whereas the present study focused on identifying a scor-
ing method that optimizes the identification of the reflex threshold,
it is possible that another scoring method optimizes the correlation
between NFR magnitude and pain intensity when using NFR as a
continuous measure of nociceptive response. Consistent with this
notion, we have shown in two separate studies that NFR Interval
Cohen’s optimizes this relationship between the reflex magnitude
and pain ratings [23,25]. While the reasons for the scoring discrep-
ancy between NFR threshold and NFR magnitude are not clear at this
time, NFR Interval Cohen’s d does produce a response that is nor-
mally distributed, a property that is not necessary for ROC analyses
but is necessary for correlational analyses.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors acknowledge Diane Turcotte for her help in scoring the
NFR threshold and Zina Trost for her comments on an earlier
draft of this manuscript.

References

because only a single bipolar AC amplifier is needed to record bi-
ceps femoris EMG activity. Thus, the NFR has much to offer to pain
researchers.

Nonetheless, to further the use of the NFR as a biomarker of no-
ceptive processes, it is important to standardize the methodology in
order to improve reliability and validity without compromising
practicality. Despite attempts by several different research groups
to generate replicable scoring criteria [1–3,8,11–15,26–29], there
has been little systematic research to compare the performance of
different criteria. The results of the present study, together with the
results of our previous report [24], indicate that NFR Interval z
score provides the best combination of overall accuracy and stability
as a standardized criterion of NFR threshold. Further, based on the
findings of the present report as well as our previous study, we rec-
ommend that researchers consider applying an NFR Interval z score
criterion of 1.4 SD in defining an NFR reflex. Adoption of a common
criterion for use with different populations, different equipment, and
in different settings would significantly enhance our ability to make
direct comparisons between studies in the future. It should be noted, however, that our efforts to identify a standardized NFR
criterion have been based on data collected from a relatively homo-
geneous population of young, healthy adults. Given that the NFR par-
cadigm is often used in research with older clinical populations, it is
advisable that further efforts be devoted to confirming our findings in
more heterogeneous samples.

Finally, whereas the present study focused on identifying a scor-
ing method that optimizes the identification of the reflex threshold,
it is possible that another scoring method optimizes the correlation
between NFR magnitude and pain intensity when using NFR as a
continuous measure of nociceptive response. Consistent with this
notion, we have shown in two separate studies that NFR Interval
Cohen’s optimizes this relationship between the reflex magnitude
and pain ratings [23,25]. While the reasons for the scoring discrep-
ancy between NFR threshold and NFR magnitude are not clear at this
time, NFR Interval Cohen’s d does produce a response that is nor-
mally distributed, a property that is not necessary for ROC analyses
but is necessary for correlational analyses.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors acknowledge Diane Turcotte for her help in scoring the
NFR threshold and Zina Trost for her comments on an earlier
draft of this manuscript.

References

Interaction between cutaneous and muscle afferent activity in polysynaptic
[2] Banic B, Petersen-Felix S, Andersen OK, Radanov BP, Villiger PM, Arendt-
Nielsen L, Curtatolo M. Evidence for spinal cord hypersensitivity in chronic pain
[3] Bourreau F, Luu M, Dobbere JF. Study of experimental pain measures and
and cingulate cortex during painful electrical nerve stimulation. Neuroreport:
Vischer TL. Neurophysiologic evidence for a central sensitization in patients
[8] Dowman R. Possible startle response contamination of the spinal nociceptive
[9] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain
[10] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain
[12] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain
[14] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain
[16] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain
[17] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain
[18] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain
[19] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain
[20] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain
[21] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain
[22] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain
[23] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain
[26] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain
[27] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain
[28] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain
[29] Edwards L, Ring C, McIntyre D, Mcintyre T. Catastrophizing is related to pain