Journal of Neuroscience Methods xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

# High inter-rater reliability in analyzing results of decomposition-based quantitative electromyography in subjects with or without neuromuscular disorder

Shaun G. Boe[a,b,*], Nathan M. Antonowicz[b,1], Vanessa W. Leung[b,2], Susan M. Shea[b,1], Toby C. Zimmerman[b,3], Timothy J. Doherty[c]

[a] School of Physiotherapy, Dalhousie University, Rm. 418, 4th Floor Forrest Building, 5869 University Avenue, Halifax, Nova Scotia, Canada B3H 3J5
[b] School of Physical Therapy, The University of Western Ontario, London, Ontario, Canada
[c] Department of Clinical Neurological Sciences, Schulich School of Medicine and Dentistry, The University of Western Ontario, London, Ontario, Canada

ARTICLE INFO

*Article history:*
Received 2 February 2010
Received in revised form 16 July 2010
Accepted 21 July 2010

*Keywords:*
Electrodiagnostic
Motor unit
Multi-center
Neuromuscular disease
Neuropathy

ABSTRACT

Decomposition-based quantitative electromyography (DQEMG) comprises a group of methods used to obtain information related to the health of the neuromuscular system. Although primarily objective, aspects of the data analysis protocol include operator decisions that may impact its reliability and reduce the applicability of the technique among multiple users. Thus, the objective of this study was to establish the inter-rater reliability of the protocol used for DQEMG analysis among five raters. Seventy data files previously obtained using DQEMG from healthy control subjects and patients with disorders of the neuromuscular system were analyzed by four novice and one experienced rater. Values obtained from this analysis were then evaluated for reliability within the novice raters and in contrast to the results of the experienced rater to examine the influence of the level of rater experience on the results obtained. The majority of the parameters associated with the number of motor unit potentials and their physiological characteristics were found to be reliable among all raters, with moderate-high ICC values observed for both the biceps brachii and first dorsal interosseous muscles. The data suggest that the level of rater experience does not greatly influence the results obtained and that the analysis can be reliably performed by a rater who is given suitable instruction. These findings are important particularly given the potential use of DQEMG as an outcome measure in multi-center studies.

© 2010 Elsevier B.V. All rights reserved.

## 1. Introduction

Decomposition-based quantitative electromyography (DQEMG) has been designed to provide information pertaining to the physiological characteristics and numbers of motor units (MUs) within a given muscle or muscle group (Doherty and Stashuk, 2003). Clinically this information is valuable in that it provides insight into the changes occurring at the level of the MU in response to disorders of the neuromuscular system. Evidence to support this has come from several studies, including the observation of decreased MU number estimates and increased MU size and complexity in patients with amyotrophic lateral sclerosis (ALS) (Boe et al., 2007) and Charcot-Marie Tooth (CMT) disease (Shy et al., 2007), in addition to demonstrating age-related MU remodeling and an associated decrease in the estimated numbers of MUs in old and very old men (McNeil et al., 2005).

To provide value as a clinical tool, it is important to ensure the data obtained using DQEMG are reliable from one test to the next both within and across raters. When studies are performed by experienced raters, DQEMG has been found to be reliable, with high

*Abbreviations:* AAR, area-to-amplitude ratio; ALS, amyotrophic lateral sclerosis; BB, biceps brachii; CMT-X, Charcot-Marie Tooth disease, Type X; COV, coefficient of variation; DQEMG, decomposition-based quantitative electromyography; EMG, electromyographic; FDI, first dorsal interosseous; FR, firing rate; ICC, intra-class correlation coefficient; IDI, interdischarge interval; MU, motor unit; MUP, motor unit potential; NpAmp, negative-peak amplitude; S-MUP, surface-detected motor unit potential.

* Corresponding author at: School of Physiotherapy, Dalhousie University, Rm. 418, 4th Floor Forrest Building, 5869 University Avenue, Halifax, Nova Scotia, Canada B3H 3J5. Tel.: +1 902 494 6360; fax: +1 902 494 1941.
*E-mail address:* s.boe@dal.ca (S.G. Boe).
[1] Address: School of Physical Therapy, *The* University of Western Ontario, Elborn College, Rm 1588, London, Ontario, Canada N6G 1H1.
[2] Address: E.W. Bickle Centre for Complex Continuing Care, Toronto Rehabilitation Institute, 130 Dunn Ave., Toronto, Ontario, Canada M6K 2R7.
[3] Address: VHA Rehab Solutions, 633 Colborne St., 2nd Floor, London, Ontario, Canada N6B 2V3.

0165-0270/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jneumeth.2010.07.024

Please cite this article in press as: Boe SG, et al. High inter-rater reliability in analyzing results of decomposition-based quantitative electromyography in subjects with or without neuromuscular disorder. J Neurosci Methods (2010), doi:10.1016/j.jneumeth.2010.07.024

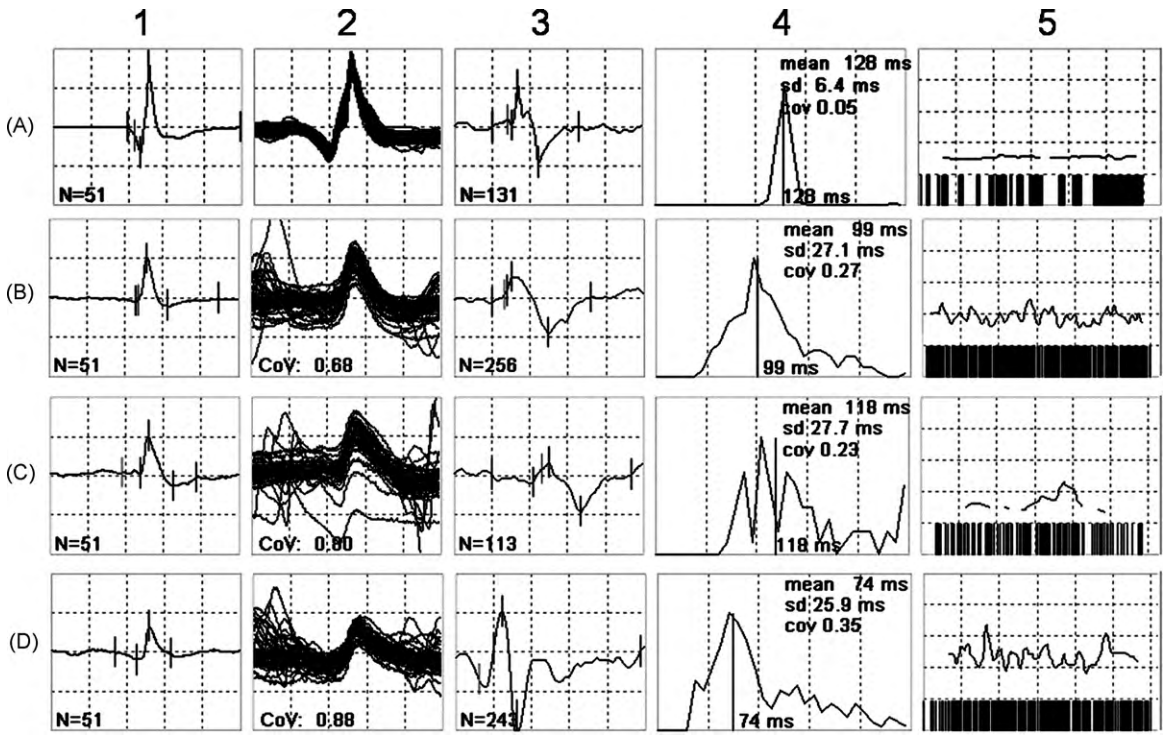


**Fig. 1.** Decomposition summary. Sample of four MUP trains (vertical, labeled A–D) that represent the decomposition of a typical needle-detected EMG signal and subsequent analysis of the needle- and surface-detected signal. Columns from left to right represent (1) prototypical needle-detected MUP; (2) individual MUPs of each MUP train superimposed in a shimmer plot; (3) surface-detected MUP and number of contributing discharges; (4) IDI histogram; and (5) firing rate vs. time plot representing MUP train discharge times and instantaneous firing rate plots. Raters are required to review each MUP train based on objective and subjective criteria to determine whether to include or exclude a particular MUP train. Briefly, this process includes: (i) does the MUP train include a minimum of 51 contributing MUPs (column 1), (ii) is the COV value associated with the IDI < 0.30 (column 4), (iii) does the firing rate vs. time plot display a consistent firing rate pattern (column 5), and (iv) does the IDI histogram have a normally distributed main peak (column 4).

intra- and inter-rater reliability observed for data collection and analysis (Boe et al., 2006, 2009; Calder et al., 2008).

Although DQEMG has been designed to utilize objective criteria during data collection and analysis, some aspects of the analysis are not amenable to quantification or classification and thus are at the discretion of the operator. These include the decisions to include or exclude needle-detected MU potential (MUP) trains and surface-detected MUPs (S-MUPs) that do not meet the objective inclusion criteria of the DQEMG software. Additionally, raters may be required to re-position inaccurately placed markers used to calculate MUP duration and amplitude measures. Although these subjective components do not impact reliability in studies performed by experienced raters (Boe et al., 2006, 2009; Calder et al., 2008), this may not be the case when analysis is performed by less experienced raters, which may occur in larger, multi-center studies. Thus, the purpose of this study was to examine the potential impact of rater experience on the intra-rater reliability of the procedures used for the analysis of DQEMG data obtained from subjects with and without neuromuscular disorder.

## 2. Methods

### 2.1. Subjects

DQEMG data from 28 healthy control subjects ($27 \pm 5$ years), nine patients with ALS ($52 \pm 12$ years) and three patients with CMT, Type X (CMT-X, $46 \pm 9$ years) who had previously undergone examination using DQEMG were analyzed in the current study. Biceps brachii (BB) data were obtained from 28 healthy control subjects and seven patients with ALS. First dorsal interosseous (FDI) data were obtained from 24 healthy control subjects, eight patients with ALS and three patients with CMT-X. All subjects had previously provided informed consent and the University of Western Ontario ethics review board approved the study.

### 2.2. DQEMG data collection

The DQEMG method, associated algorithms and data collection protocols for the FDI and BB muscles have been described in detail elsewhere (Boe et al., 2004, 2007; Doherty and Stashuk, 2003; Stashuk, 1999). Utilizing a series of pattern recognition algorithms in addition to spike-triggered averaging, DQEMG is able to breakdown both a needle and surface-detected EMG signal, acquired simultaneously during a voluntary muscle contraction, into their individual needle and surface-detected MUPs (Stashuk, 1999, 2001). Briefly, DQEMG decomposes the composite needle-detected EMG signal into its constituent MUP trains using shape and temporal information related to the individual MUP discharges in addition to MU firing time statistics. Using these needle-detected MUPs as triggers for spike-triggered averaging, a sample of MUPs, detected via surface electrodes (S-MUPs), are obtained. These S-MUPs are representative of the sizes of the MUs in the underlying muscle of interest (Stalberg, 1980; Stalberg and Fawcett, 1982). Parameters associated with the needle and surface-detected MUPs provide information regarding MU size, complexity and discharge rate. Although not presented here, if a representative sample of these S-MUPs is attained ($\geq$20), a mean S-MUP size can be determined and divided into a corresponding size-related parameter of a maximal M wave to produce a MU number estimate (Boe et al., 2004).

### 2.3. Raters

The novice raters were four second-year students enrolled in a Masters level clinical program at the University of Western Ontario.

Please cite this article in press as: Boe SG, et al. High inter-rater reliability in analyzing results of decomposition-based quantitative electromyography in subjects with or without neuromuscular disorder. J Neurosci Methods (2010), doi:10.1016/j.jneumeth.2010.07.024

Each had previously demonstrated a fundamental understanding of neurophysiology, completing both undergraduate and graduate level courses in the neurosciences. The novice raters had no previous experience with either qualitative or quantitative EMG (including DQEMG). To investigate the effect of rater experience on data analysis, the results of the novice group were compared to those of a rater who has used DQEMG extensively for ~6 years (i.e., the 'experienced rater').

### 2.4. Data analysis

Prior to data analysis, the novice raters were oriented to the DQEMG software and quantitative EMG examination during three group instructional sessions with the experienced rater. During the first session, the expert rater described in detail and demonstrated to the novice raters the process of DQEMG data analysis. Data analysis primarily involves the inclusion or exclusion of MUP trains and the re-positioning of inaccurately placed markers following the automated portion of the analysis. Inclusion or exclusion of the MUP trains is based on both objective and subjective criteria, and involves the rater addressing four sequential points, which are highlighted in Fig. 1 and outlined in the following text: (1) does the MUP train include a minimum of 51 contributing MUPs; (2) is the coefficient of variation (COV) value associated with the inter-discharge interval (IDI) less than 0.30 (Fuglevand et al., 1993); (3) does the firing rate vs. time plot display a consistent firing rate pattern; and (4) does the IDI histogram have a normally (Gaussian) distributed main peak. To illustrate this process, a sample of MUP trains is presented as an example (Fig. 1). Based on the objective criteria of points 1 and 2, MUP trains 'A, B and C' would be included in the analysis, whereas MUP train 'D' (which represents a poor example) would be excluded due to a COV value in excess of 0.30 (Fig. 1, column 4). Steps 3 and 4 comprise the subjective component of the analysis, requiring a review and decision with regard to suitability for acceptance. In comparison to MUP train 'A', which represents a prototypical case for inclusion, MUP trains 'B' and 'C' contain features that complicate the decision making process. For instance, although MUP train 'B' displays a relatively normally distributed IDI histogram main peak (Fig. 1, column 4), there is a degree of variability in the firing rate vs. time plot (Fig. 1, column 5). There is less variability in the firing rate vs. time plot for MUP train 'C' but the IDI histogram has multiple peaks (Fig. 1, columns 5 and 4, respectively). For this sample, MUP trains 'A' and 'B' would be included in the analysis, with MUP train 'C' ultimately being excluded due to the variability observed in the IDI histogram.

Subsequent to the inclusion/exclusion process, the rater reviews the automated marker placements for the needle- and surface-detected MUPs associated with the remaining MUP trains. This process includes examination of individual MUPs in a magnified view (i.e., decreased amplitude and duration scaling) to ensure accurate placement of markers at the onset, negative peak, positive peak and end of the MUP waveform (Fig. 2). Markers may be manually re-positioned if they are deemed by the rater to be incorrectly placed. Lastly, visual inspection of the negative onset of the S-MUP waveform is performed at this time to ensure it occurred within 10 ms of the onset of the needle-detected MUP. Surface-detected MUPs that do not fit this criterion are excluded from further analysis.

At the conclusion of the first session, each of the novice raters was provided with the same set of six DQEMG data files obtained from the FDI ($n = 3$) and BB muscles ($n = 3$). Each set included both control and patient data. The novice raters were instructed to familiarize themselves with the DQEMG software and to practice analyzing the six data files using the steps described above. A second session occurred one week later to review and address any questions regarding the data analysis. During the final ses-

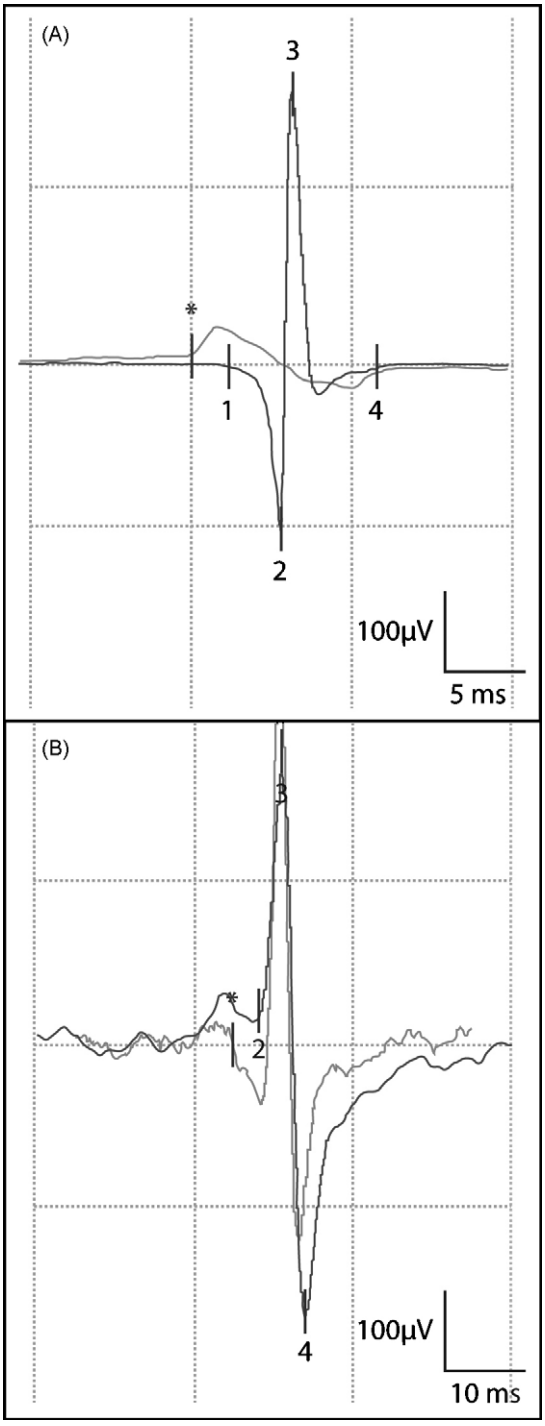


**Fig. 2.** (A) Needle-detected MUP (darker tracing) superimposed on its representative surface-detected MUP (lighter tracing). Markers 1-4 represent: (1) onset – point of deviation from baseline (negative or positive); (2) positive peak – point of greatest positive deviation from the baseline; (3) negative peak – point of greatest negative deviation from the baseline; (4) end – point of cessation of EMG activity (i.e., a return to baseline with no subsequent negative or positive deviation). * represents the onset of the surface-detected MUP. (B) Surface-detected MUP (darker tracing) superimposed on its representative needle-detected MUP (lighter tracing). Markers 2–3 represent: (2) negative onset – point of greatest negative deviation (represented by highest slope) from the baseline; (3) negative peak – point of greatest negative deviation from the baseline; (4) positive peak – point of greatest positive deviation from the baseline. * represents the onset of the needle-detected MUP.

Please cite this article in press as: Boe SG, et al. High inter-rater reliability in analyzing results of decomposition-based quantitative electromyography in subjects with or without neuromuscular disorder. J Neurosci Methods (2010), doi:10.1016/j.jneumeth.2010.07.024

**Table 1**
Biceps brachii values across five raters. Values are expressed as mean ± SD representing results of data analysis for 28 BB data files for each rater. NpAmp, negative-peak amplitude; P–P Vol, peak–peak voltage. See text for additional abbreviations.

| Parameter | Rater 1 | Rater 2 | Rater 3 | Rater 4 | Rater 5[a] | ICC value | |
|---|---|---|---|---|---|---|---|
| | | | | | | 4 Raters | 5 Raters |
| Needle MUP duration (ms) | 10.0 ± 1.8 | 9.9 ± 1.7 | 9.5 ± 1.6 | 9.3 ± 1.6 | 9.3 ± 1.5 | 0.847 | 0.776 |
| AAR | 1.3 ± 0.2 | 1.3 ± 0.2 | 1.4 ± 0.2 | 1.3 ± 0.2 | 1.4 ± 0.2 | 0.916 | 0.862 |
| Number of phases | 2.5 ± 0.4 | 2.5 ± 0.5 | 2.4 ± 0.3 | 2.4 ± 0.4 | 2.6 ± 0.2 | 0.379 | 0.299 |
| Number of turns | 3.0 ± 0.4 | 3.0 ± 0.4 | 2.8 ± 0.4 | 3.0 ± 0.4 | 2.7 ± 0.4 | 0.819 | 0.735 |
| Firing rate (Hz) | 12.0 ± 1.2 | 12.0 ± 1.3 | 12.1 ± 1.2 | 11.9 ± 1.2 | 12.2 ± 1.2 | 0.970 | 0.957 |
| COV IDI histogram | 0.1 ± 0.01 | 0.1 ± 0.01 | 0.1 ± 0.01 | 0.1 ± 0.01 | 0.08 ± 0.01 | 0.901 | 0.465 |
| Surface-detected NpAmp (μV) | 60.1 ± 21.3 | 60.4 ± 19.8 | 56.0 ± 20.1 | 61.7 ± 20.5 | 55.5 ± 19.5 | 0.962 | 0.950 |
| Number of MUPs | – | 27 ± 7 | – | 28 ± 7 | 26 ± 7 | 0.964[b] | 0.716[b] |
| Needle P–P Vol (μV) | – | 386 ± 63.5 | – | 387 ± 65.3 | 348 ± 65.8 | 0.954[b] | 0.831[b] |

[a] Experienced rater.
[b] ICC values representative of 2 and 3 raters, respectively.

sion (which occurred one day following the preceding session), any additional questions were addressed and each of the raters received the data files ($N = 70$) for analysis. Raters were instructed to continue performing the analysis in the periods between training sessions to generate additional questions and to implement discussion points from the previous session. Throughout the training period, each rater spent approximately 12 h practicing data analysis (4 h per week) and each was provided with equal opportunity to discuss any questions or concerns with the experienced rater. Questions posed by an individual rater and the subsequent response were presented to the group to ensure consistency in the information delivered to the novice raters throughout the training period.

### 2.5. Statistics

Prior to statistical analysis, a mean value for the size, complexity and number of needle-detected MUPs, and the size of the S-MUPs was obtained for BB and FDI data for each individual rater. Data were then tested for normality, and those that were outside the normal distribution were analyzed using non-parametric statistics. Thus, comparison of these mean values among the four novice raters was performed using either a standard one-way analysis of variance or a Kruskal–Wallis one-way analysis of variance, with an a priori alpha level of $p < 0.05$ denoting significance (GraphPad Prism 4; GraphPad Software v. 4.02, San Diego, CA). If significant differences were detected, post hoc analyses were performed to determine the significant interactions using either the Tukey or Dunn test. Reliability amongst the novice raters was then determined using a two-way random, single measure intraclass correlation coefficient (ICC; SPSS v.15.0, Chicago, IL). This particular form of ICC was used to determine reliability as it accounts for the consistency of the values and their absolute agreement and is useful in assessing the generalizability of the values from each of the raters in the current study to future operators (Bartko, 1966; Laschinger, 1992; Muller and Buttner, 1994; Shrout and Fleiss, 1979). These same analyses were repeated following the addition of the results from the experienced rater to allow for contrast with those of the novice group. Lastly, to allow for comparison to a recently published study, two additional parameters (MUP peak–peak voltage and the total number of MUPs included post-data analysis) were analyzed for each data file in a sub-set of the raters (two novice and the experienced rater). As a portion of this secondary analysis was performed between only the two members of the novice group, either a standard pairwise *t* test or the Wilcoxon ranked sums test was employed (SPSS v.15.0 Graduate Student Package, Chicago, IL). Patient and control data were analyzed separately to allow for assessment of reliability among data obtained from individuals with and without neuromuscular disorder. All data are presented as mean values ± standard deviation.

## 3. Results

### 3.1. Biceps brachii – controls and patients

#### 3.1.1. Novice raters

Mean values among the novice raters were similar for both control and patient data files. Parameters associated with the size, complexity and discharge rate of the needle-detected MUPs (i.e., duration, area-to-amplitude ratio (AAR), numbers of phases and turns and firing rate) were examined across the four novice raters. With the exception of MUP duration ($p < 0.05$), all other needle parameters were found to be similar ($p > 0.05$, Tables 1 and 3). Post hoc analysis of the MUP duration data revealed differences amongst three of the four novice raters (Tables 1 and 3). Values associated with the COV of the IDI histogram and the surface-detected MUP size (based on negative-peak amplitude), were also found to be similar across the novice raters ($p > 0.05$). Lastly, analysis within a sub-set of the novice raters (raters 2 and 3, Tables 1 and 3) revealed no differences ($p > 0.05$) for either the number of MUP trains included in the analysis or the amplitude of the needle-detected MUPs (based on peak–peak voltage). Consistent with the aforementioned results, high ICC values were observed for all parameters across the four raters for both control and patient data, with the exception of the number of phases which had a low ICC value in both subject groups (Tables 1 and 3).

#### 3.1.2. Novice and experienced raters

In general, the inclusion of the experienced rater (rater 5) in the statistical analyses increased the variability of the data as compared to within the novice raters alone. Motor unit potential duration, numbers of phases and turns, and the COV values associated with the IDI were all found to differ significantly with the addition of the experienced rater. Of these parameters, MUP duration and the COV values were different for both control and patient data, whereas the numbers of phases and turns differed only within the control subjects (Tables 1 and 3). Post hoc analyses revealed differences between one of the novice raters and the experienced rater for MUP duration and the numbers of phases and turns (Tables 1 and 3); however, differences between each of the novice raters and the experienced rater were noted for the COV values (Tables 1 and 3). No significant differences were observed for AAR, firing rate or S-MUP negative-peak amplitude across the five raters for either the control or patient data. The number of MUP trains included in the analysis and needle-detected MUP peak–peak voltage differed amongst the sub-set of novice raters and the experienced rater for the control and patient data, with the exception of the

Please cite this article in press as: Boe SG, et al. High inter-rater reliability in analyzing results of decomposition-based quantitative electromyography in subjects with or without neuromuscular disorder. J Neurosci Methods (2010), doi:10.1016/j.jneumeth.2010.07.024

**Table 2**
First dorsal interosseous values across five raters. Values are expressed as mean ± SD representing results of data analysis for 24 FDI data files for each rater. NpAmp, negative-peak amplitude; P–P Vol, peak–peak voltage. See text for additional abbreviations.

| Parameter | Rater 1 | Rater 2 | Rater 3 | Rater 4 | Rater 5[a] | ICC value | |
|---|---|---|---|---|---|---|---|
| | | | | | | 4 Raters | 5 Raters |
| Needle MUP duration (ms) | 9.5 ± 1.5 | 9.7 ± 1.3 | 9.2 ± 1.5 | 7.9 ± 1.3 | 7.9 ± 1.4 | 0.543 | 0.521 |
| AAR | 1.3 ± 0.2 | 1.3 ± 0.2 | 1.3 ± 0.2 | 1.3 ± 0.2 | 1.3 ± 0.2 | 0.905 | 0.888 |
| Number of phases | 2.7 ± 0.3 | 2.8 ± 0.5 | 2.5 ± 0.3 | 2.6 ± 0.3 | 2.5 ± 0.3 | 0.319 | 0.243 |
| Number of turns | 3.3 ± 0.4 | 3.5 ± 1.0 | 3.2 ± 0.3 | 3.2 ± 0.3 | 2.9 ± 0.3 | 0.293 | 0.259 |
| Firing Rate (Hz) | 12.4 ± 2.7 | 13.0 ± 2.1 | 12.7 ± 1.4 | 12.7 ± 1.5 | 13.0 ± 1.4 | 0.608 | 0.650 |
| COV IDI histogram | 0.13 ± 0.01 | 0.13 ± 0.02 | 0.13 ± 0.02 | 0.14 ± 0.02 | 0.11 ± 0.02 | 0.902 | 0.651 |
| Surface-detected NpAmp (μV) | 135 ± 47.3 | 135 ± 48.1 | 129 ± 44.5 | 135 ± 45.3 | 131 ± 46.2 | 0.971 | 0.968 |
| Number of MUPs | – | 29 ± 6 | – | 30 ± 7 | 35 ± 11 | 0.853[b] | 0.605[b] |
| Needle P–P Vol (μV) | – | 516.1 ± 108.0 | – | 500.1 ± 100.0 | 450.0 ± 96.2 | 0.884[b] | 0.760[b] |

[a] Experienced rater.
[b] ICC values representative of 2 and 3 raters, respectively.

number of MUP trains included for the control group. Post hoc analyses revealed differences between each of the novice raters and the experienced rater. Although a number of the parameters investigated demonstrated differences between the experienced and at least one of the novice raters, ICC values ranged from moderate-high for the majority of the parameters. Similar to the findings amongst the novice raters, the ICC values for the five raters were lowest for the number of phases in the control and patient data (Tables 1 and 3).

### 3.2. First dorsal interosseous – controls and patients

#### 3.2.1. Novice raters

Although similar values were observed for MUP duration and the number of phases, statistical analysis revealed significant differences among the four novice raters for the control and patient data (Tables 2 and 4). Post hoc analyses revealed differences between raters 1 and 4 for MUP duration (controls and patients), and raters 2 and 3 (controls), and 1 and 3 (patients), for the number of phases. With the exception of the COV values associated with the IDI histogram of the patient data ($p < 0.05$; with post hoc results revealing a difference between rater 1 compared to 4), all other parameters including number of turns, AAR, COV of the IDI (controls), firing rate and S-MUP negative-peak amplitude, were similar ($p > 0.05$) for the control and patient data. Examination of the two additional parameters (number of MUPs included post-data analysis and needle-detected MUP peak–peak voltage) revealed no significant differences between the sub-set of novice raters (2 and 4) for either the control or patient data. Analysis performed using the ICC to assess reliability generally paralleled the above highlighted statistical pattern, with moderate to high values observed for parameters that were not statistically different (Tables 2 and 4). An exception to this trend was noted for the numbers of turns, as a low ICC value was observed despite the non-significant finding.

#### 3.2.2. Novice and experienced raters

Similar to the BB, analyses of the FDI data which included the experienced rater resulted in significant differences for a number of the parameters examined for control and patient data. For both data sets, significant differences were found across the five raters for the number of phases, MUP duration and the COV values associated with the IDI. Post hoc analyses revealed differences between one of the novice raters (rater 2) and the experienced rater for the number of phases, with differences noted between each of the novice raters and the experienced rater for MUP duration and the COV values. For the remaining parameters (numbers of turns, AAR, firing rate and S-MUP negative-peak amplitude) no significant differences were observed for the control or patient data, with the exception of the number of turns for the control subject data. Post hoc analysis of the number of turns revealed differences between raters 1, 2 and 3 when compared to rater 5 (Table 2). Examination of the number of MUPs included post-data analysis and of the needle-detected MUP peak–peak voltage in the sub-set of novice raters compared to the experienced rater revealed significant differences across control and patient data for both parameters (Tables 2 and 4). The exception was peak–peak voltage, which was similar across the raters for the patient data (Tables 2 and 4). Where significant differences were detected, post hoc analyses revealed differences between each of novice raters (raters 2 and 4) and the experienced rater.

## 4. Discussion

The goal of this study was to examine the inter-rater reliability of DQEMG data analysis, with a focus on the level of rater experience on the reliability of the results obtained. The observation of moderate-high levels of reliability for the majority of the parameters examined suggests that following a brief training period, novice raters are able to independently achieve similar results, both across their group and in comparison to an experienced rater. Although moderate-high levels of reliability were observed, mean values for some parameters were found to be significantly different amongst the raters. Potential sources of this variability and their influence on the clinical application of DQEMG are discussed below.

### 4.1. Inter-rater differences

Differences observed among raters for the parameters reported in the present study can be attributed to two sources of variability in the data analysis process. First, the presence of a bias in the type of MUP trains that are included or excluded by a given rater can lead to differences across raters in the values associated with MU size. Second, variability in the placement of markers denoting MUP onset and end will result in differences in the values dependent on these markers for quantification (i.e., duration, numbers of phases and turns).

#### 4.1.1. Inclusion and exclusion of MUP trains

As highlighted in Section 2, raters make decisions regarding the inclusion or exclusion of MUP trains based on objective and subjective criteria. Due to the subjective component of this process, it is possible that different raters may include/exclude MUP trains with a bias to MUs at the extremes of the size spectrum (i.e., smaller or larger MUs) resulting in MUP size-related values that are divergent from the other raters. For instance, in the example provided (Fig. 1), MUP trains 'C' and 'D' are considered 'merged trains', or trains that represent the MUPs of two MUs firing synchronously or in close temporal succession (Stashuk, 2001, 1999). These 'merged trains' represent temporal and morphological information reflective of the

Please cite this article in press as: Boe SG, et al. High inter-rater reliability in analyzing results of decomposition-based quantitative electromyography in subjects with or without neuromuscular disorder. J Neurosci Methods (2010), doi:10.1016/j.jneumeth.2010.07.024

**Table 3**
Biceps brachii values across five raters (patients). Values are expressed as mean ± SD representing results of data analysis for 7 BB data files for each rater. NpAmp, negative-peak amplitude; P–P Vol, peak–peak voltage. See text for additional abbreviations.

| Parameter | Rater 1 | Rater 2 | Rater 3 | Rater 4 | Rater 5[a] | ICC value | |
|---|---|---|---|---|---|---|---|
| | | | | | | 4 Raters | 5 Raters |
| Needle MUP duration (ms) | 12.7 ± 2.6 | 13.2 ± 2.4 | 12.3 ± 1.9 | 11.5 ± 2.0 | 11.9 ± 1.7 | 0.767 | 0.777 |
| AAR | 1.8 ± 0.3 | 1.8 ± 0.3 | 1.8 ± 0.3 | 1.8 ± 0.3 | 1.8 ± 0.3 | 0.920 | 0.924 |
| Number of phases | 2.9 ± 0.5 | 2.9 ± 0.4 | 3.1 ± 0.7 | 2.7 ± 0.2 | 2.9 ± 0.2 | −0.060 | −0.022 |
| Number of turns | 3.5 ± 0.8 | 3.6 ± 0.8 | 3.3 ± 0.8 | 3.5 ± 0.8 | 3.2 ± 0.5 | 0.953 | 0.881 |
| Firing rate (Hz) | 12.3 ± 1.8 | 12.2 ± 1.7 | 12.3 ± 1.7 | 12.2 ± 1.8 | 12.2 ± 1.4 | 0.987 | 0.973 |
| COV IDI histogram | 0.12 ± 0.03 | 0.13 ± 0.03 | 0.13 ± 0.03 | 0.13 ± 0.03 | 0.11 ± 0.03 | 0.955 | 0.867 |
| Surface-detected NpAmp (μV) | 52.3 ± 16.1 | 55.7 ± 24.2 | 53.0 ± 25.8 | 57.7 ± 24.0 | 58.0 ± 34.0 | 0.911 | 0.893 |
| Number of MUPs | – | 15 ± 7 | – | 15 ± 6 | 21 ± 7 | 0.974[b] | 0.673[b] |
| Needle P–P Vol (μV) | – | 501 ± 92.1 | – | 501 ± 90.6 | 460 ± 95.2 | 0.955[b] | 0.924[b] |

[a] Experienced rater.
[b] ICC values representative of 2 and 3 raters, respectively.

aggregate activity of two MUs and they often represent MUs that contribute significantly to the overall EMG and have large amplitude MUPs (both needle- and surface-detected). A bias to their inclusion considerably increases the resultant MUP size-related values (unpublished observation). An additional consideration with respect to MUP inclusion/exclusion is the number of MUPs each rater includes post-analysis. Variability in this value would indicate differences in the exclusion rate across raters, as well as suggest a difference in the criteria used in the inclusion/exclusion process.

While it is likely that the aforementioned factors associated with the inclusion/exclusion process contribute a degree of variability, based on our results it is reasonable to suggest that their contribution is small relative to the process of marker placement (see next section for detailed discussion). Amongst the novice raters, the observation of similar IDI histogram COV values for controls and patients indicate that the raters were managing the inclusion/exclusion of these 'merged trains' in a similar manner. Additionally, the observation of low COV values among the novice raters (Tables 1–4) indicates that they were effectively excluding 'merged trains', as elevated COV values would have resulted from their inclusion in the analysis. Comparison of the novice group with the experienced rater however revealed significantly lower COV values for the experienced rater for the control and patient data in both muscles examined. While this difference may be interpreted as a tendency for the experienced rater to exclude a greater number of 'merged trains' in contrast to the novice raters, the COV values observed do not support this. Although the experienced rater had lower COV values for both muscles and across both subject groups, the largest difference in COV values between a novice and the experienced rater was 0.03 (0.11 vs. 0.14; Table 4). More importantly, the COV values noted for all raters were well below the 0.30 cut-off, substantiating the conclusion that this group was able to appropriately exclude those MUP trains that were not representative of single MU discharges.

Data pertaining to MU size further support a parallel approach to MUP inclusion/exclusion between the novice group and experienced rater. Had differences been present in the types of MUP trains included or excluded, it would be reasonable to expect differences in the size-related values of the MUs resulting from the analysis. This was not the case however as S-MUP size, which is reflective of MU size (Stalberg, 1980; Stalberg and Fawcett, 1982), was found to be similar across all raters for both muscles and control and patient data (Tables 1–4). In-line with this finding were considerably high S-MUP reliability values across the five raters, with ICC values ranging from 0.893 to 0.971 (Tables 1–4).

Examination of the number of MUPs included in the analysis was limited to two novice (raters 2 and 4) and the experienced rater. These data, while limited, confirm the novice raters had a similar rate of MUP train exclusion. In comparison to the experienced rater however, the data revealed a trend for the experienced rater to include a greater number of MUP trains, particularly for the patient data (Tables 3 and 4). This finding does not seem to reflect the inclusion of a greater number of 'merged trains' by the experienced rater, as outlined above, nor does it seem to impact on the size of the MUs being included in the analysis, based on the similarities observed for S-MUP negative-peak amplitude. It does indicate that the novice raters were slightly more conservative in accepting MUP trains, particularly those acquired from patient populations. It could be speculated that this finding was the result of an increased level of difficulty in analyzing data obtained from individuals with neuromuscular disease. Specifically, our experience with this type of data suggests the MUP trains have greater variability, likely resulting from an ongoing process of denervation–reinnervation, and its associated motor control challenges. Increased variability in the MUP train data in-turn increases the difficulty associated with

**Table 4**
First dorsal interosseous values across five raters (patients). Values are expressed as mean ± SD representing results of data analysis for 11 FDI data files for each rater. NpAmp, negative-peak amplitude; P–P Vol, peak–peak voltage. See text for additional abbreviations.

| Parameter | Rater 1 | Rater 2 | Rater 3 | Rater 4 | Rater 5[a] | ICC values | |
|---|---|---|---|---|---|---|---|
| | | | | | | 4 Raters | 5 Raters |
| Needle MUP duration (ms) | 13.3 ± 2.6 | 13.8 ± 2.4 | 12.7 ± 2.1 | 11.3 ± 2.4 | 11.9 ± 2.2 | 0.650 | 0.664 |
| AAR | 1.9 ± 0.4 | 1.9 ± 0.4 | 1.9 ± 0.4 | 1.9 ± 0.4 | 1.9 ± 0.4 | 0.953 | 0.953 |
| Number of phases | 2.9 ± 0.3 | 2.9 ± 0.3 | 2.7 ± 0.2 | 2.7 ± 0.3 | 2.8 ± 0.2 | 0.470 | 0.471 |
| Number of turns | 3.7 ± 0.5 | 3.6 ± 0.5 | 3.4 ± 0.6 | 3.4 ± 0.4 | 3.2 ± 0.4 | 0.721 | 0.670 |
| Firing rate (Hz) | 11.1 ± 2.8 | 10.9 ± 2.8 | 11.1 ± 2.8 | 11.0 ± 2.9 | 11.5 ± 2.7 | 0.987 | 0.978 |
| COV IDI histogram | 0.13 ± 0.02 | 0.14 ± 0.02 | 0.14 ± 0.02 | 0.14 ± 0.02 | 0.11 ± 0.02 | 0.843 | 0.686 |
| Surface-detected NpAmp (μV) | 165 ± 91.0 | 169 ± 88.3 | 150 ± 79.2 | 173 ± 90.5 | 157 ± 74.0 | 0.910 | .0911 |
| Number of MUPs | – | 18 ± 6 | – | 18 ± 6 | 22 ± 6 | 0.895[b] | 0.730[b] |
| Needle P–P Vol (μV) | – | 1136 ± 602.3 | – | 1139 ± 531.2 | 1083 ± 706.6 | 0.957[b] | 0.939[b] |

[a] Experienced rater.
[b] ICC values representative of 2 and 3 raters, respectively.

Please cite this article in press as: Boe SG, et al. High inter-rater reliability in analyzing results of decomposition-based quantitative electromyography in subjects with or without neuromuscular disorder. J Neurosci Methods (2010), doi:10.1016/j.jneumeth.2010.07.024

making subjective decisions regarding inclusion and exclusion, and it appears the novice raters 'erred on the side of caution', which resulted in the inclusion of fewer MUP trains relative to the experienced rater. Unfortunately, the present data set does not allow for investigation into why the novice raters have a slightly higher rate of MUP train exclusion, and thus this finding may warrant additional study in the future.

#### 4.1.2. Marker placement

Although a majority of the parameters reported demonstrated moderate-high reliability based on the ICC, comparison of values representing MU size (MUP duration) and complexity (number of turns and phases) revealed significant differences amongst this group of raters. This is of particular concern, as this variability potentially impacts on the ability to discern differences from one test to the next between those changes that result from disease progression or treatment and those that result from variability inherent in the analysis procedure. While the degree of variability may be influenced by the process of MUP train inclusion/exclusion as highlighted above, the current results, coupled with previous findings (Calder et al., 2008), indicate the primary source of variability is in the process of marker placement on the needle-detected MUPs. Placement of these markers, in particular the onset and end markers, provide the limits from which MUP duration and the number of phases and turns are calculated. Due to inconsistencies in the automated process of marker placement (Bischoff et al., 1994; Bromberg et al., 1999; Stalberg et al., 1986; Takehara et al., 2004), manual review and adjustment of markers is routinely performed in DQEMG data analysis (see Section 2 and Fig. 2 for details). Our findings indicate that these manual adjustments are not performed reliably across raters, as significant differences for MUP duration were detected for both muscles across all raters for the control and patient data. Consistent with this finding, low ICC values were observed for the number of phases and turns. This may be related to characteristics inherent in the statistical analyses, as the ICC is strongly influenced by the variance of the parameter or variable in the population which is being assessed (Bartko, 1966; Shrout and Fleiss, 1979). For example, when the ICC is applied to a variable within a population that is near-uniform (i.e., displays little variation), discrepancies amongst raters tend to be magnified, resulting in the generation of a lower ICC value than what would have been produced if it was applied to a population with a greater degree of variability. This fact notwithstanding, the finding of significant differences for MUP duration and the number of phases and turns between raters is similar to previous studies examining intra- and inter-rater reliability among experienced users of quantitative electromyography (Boe et al., 2006, 2009; Calder et al., 2008; Rodriguez et al., 2007).

Consistent with the conclusions of Calder et al. (2008), the variability associated with measures of MUP duration observed with DQEMG may have a significant impact on the study of clinical populations, as this measure is often used in the identification of neuropathic and myopathic disorders (Bischoff et al., 1994; Nandedkar et al., 1988; Stewart et al., 1989). This potential clinical limitation of DQEMG (and other clinically viable quantitative EMG systems) stresses that an alternative means of assessing muscle and nerve disease using quantitative data generated by DQEMG would be advantageous. Previous studies have identified AAR as a useful index of MU size or 'thickness' that is of similar utility to MUP duration in assessing neuropathy and myopathy (Nandedkar et al., 1988). Area-to-amplitude ratio has also been reported to be robust to poor signal-to-noise ratio, which is a contributing factor to inaccurate marker placement (Nandedkar et al., 1988). In the present study, AAR was remarkably consistent across the group of raters, with no differences detected across all raters for control or patient data for either muscle examined. Reliability amongst all raters was also high (Tables 1–4), a finding which is consistent with previous intra- and inter-rater studies of experienced operators (Boe et al., 2009; Calder et al., 2008). These findings suggest that it may be preferable to utilize AAR rather than MUP duration to identify neuropathic and myopathic disorders.

### 4.2. Summary and conclusions

Our results indicate that the majority of the quantitative MU analysis parameters can be reliably analyzed across a group of novice raters, and that these results are similar to those generated by an experienced rater. It is important to acknowledge however that while our index of reliability was often high, significant differences were detected amongst the raters for some of the values measured. Although this variability does not appreciably impact the results obtained for many parameters, the clinical utility of other parameters, including MUP duration, are influenced by the degree of variability. As such, future work should be directed towards improving the automated process of marker placement and the procedure used for manual adjustments, with the goal of improving the reliability of these measures and in-turn their clinical utility. With regard to the use of DQEMG in multi-center studies however, it may be important to consider the need of having a single examiner review individual data sets to ensure consistency in the analysis procedure.

Overall, DQEMG has the potential to be a viable quantitative tool to monitor changes at the level of the MU in individuals with or without neuromuscular disorder, further supporting its use in multi-center studies.

## References

Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966;19:3–11.
Bischoff C, Stalberg E, Falck B, Eeg-Olofsson KE. Reference values of motor unit action potentials obtained with multi-MUAP analysis. Muscle Nerve 1994;17:842–51.
Boe SG, Dalton BH, Harwood B, Doherty TJ, Rice CL. Inter-rater reliability of motor unit number estimates and quantitative motor unit analysis in the tibialis anterior muscle. Clin Neurophysiol 2009;120:947–52.
Boe SG, Stashuk DW, Doherty TJ. Motor unit number estimates and quantitative motor unit analysis in healthy subjects and patients with amyotrophic lateral sclerosis. Muscle Nerve 2007;36:62–70.
Boe SG, Stashuk DW, Doherty TJ. Motor unit number estimation by decomposition-enhanced spike-triggered averaging: control data, test–retest reliability, and contractile level effects. Muscle Nerve 2004;29:693–9.
Boe SG, Stashuk DW, Doherty TJ. Within-subject reliability of motor unit number estimates and quantitative motor unit analysis in a distal and proximal upper limb muscle. Clin Neurophysiol 2006;117:596–603.
Bromberg MB, Smith AG, Bauerle J. A comparison of two commercial quantitative electromyographic algorithms with manual analysis. Muscle Nerve 1999;22:1244–8.
Calder KM, Agnew MJ, Stashuk DW, McLean L. Reliability of quantitative EMG analysis of the extensor carpi radialis muscle. J Neurosci Methods 2008;168:483–93.
Doherty TJ, Stashuk DW. Decomposition-based quantitative electromyography: methods and initial normative data in five muscles. Muscle Nerve 2003;28:204–11.
Fuglevand AJ, Winter DA, Patla AE. Models of recruitment and rate coding organization in motor-unit pools. J Neurophysiol 1993;70:2470–88.
Laschinger HK. Intraclass correlations as estimates of interrater reliability in nursing research. West J Nurs Res 1992;14:246–51.
McNeil CJ, Doherty TJ, Stashuk DW, Rice CL. Motor unit number estimates in the tibialis anterior muscle of young, old, and very old men. Muscle Nerve 2005;31:461–7.
Muller R, Buttner P. A critical discussion of intraclass correlation coefficients. Stat Med 1994;13:2465–76.
Nandedkar SD, Barkhaus PE, Sanders DB, Stalberg EV. Analysis of amplitude and area of concentric needle EMG motor unit action potentials. Electroencephalogr Clin Neurophysiol 1988;69:561–7.
Rodriguez I, Gila L, Malanda A, Gurtubay IG, Mallor F, Gomez S, et al. Motor unit action potential duration, I: variability of manual and automatic measurements. J Clin Neurophysiol 2007;24:52–8.
Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.
Shy ME, Siskind C, Swan ER, Krajewski KM, Doherty T, Fuerst DR, et al. CMT1X phenotypes represent loss of GJB1 gene function. Neurology 2007;68:849–55.
Stalberg E. Macro EMG, a new recording technique. J Neurol Neurosurg Psychiatr 1980;43:475–82.

Please cite this article in press as: Boe SG, et al. High inter-rater reliability in analyzing results of decomposition-based quantitative electromyography in subjects with or without neuromuscular disorder. J Neurosci Methods (2010), doi:10.1016/j.jneumeth.2010.07.024

Stalberg E, Andreassen S, Falck B, Lang H, Rosenfalck A, Trojaborg W. Quantitative analysis of individual motor unit potentials: a proposition for standardized terminology and criteria for measurement. J Clin Neurophysiol 1986;3:313–48.

Stalberg E, Fawcett PR. Macro EMG in healthy subjects of different ages. J Neurol Neurosurg Psychiatr 1982;45:870–8.

Stashuk D. EMG signal decomposition: how can it be accomplished and used? J Electromyogr Kinesiol 2001;11:151–73.

Stashuk DW. Decomposition and quantitative analysis of clinical electromyographic signals. Med Eng Phys 1999;21:389–404.

Stewart CR, Nandedkar SD, Massey JM, Gilchrist JM, Barkhaus PE, Sanders DB. Evaluation of an automatic method of measuring features of motor unit action potentials. Muscle Nerve 1989;12:141–8.

Takehara I, Chu J, Schwartz I, Aye HH. Motor unit action potential (MUAP) parameters affected by editing duration cursors. Electromyogr Clin Neurophysiol 2004;44:265–9.

Please cite this article in press as: Boe SG, et al. High inter-rater reliability in analyzing results of decomposition-based quantitative electromyography in subjects with or without neuromuscular disorder. J Neurosci Methods (2010), doi:10.1016/j.jneumeth.2010.07.024