Performance of Individual Items of the Quantitative Myasthenia Gravis Score

Carolina Barnett MD\textsuperscript{1,2}, Vera Bril MD, FRCP\textsuperscript{1,2} and Aileen M. Davis MSc, PhD\textsuperscript{3,4}

1. Department of Medicine (Neurology), University of Toronto
2. Toronto General Hospital, University Health Network
3. Health Care and Outcomes Research, Toronto Western Research Institute, University Health Network
4. Institute of Health Policy, Management and Evaluation, University of Toronto

Corresponding author: Carolina Barnett 5EC-309, Toronto General Hospital, 200 Elizabeth St, Toronto, ON, Canada, M5G 2C4
Phone: 416-340-3315 Fax: 416-340-4189 Email: c.barnetttapia@utoronto.ca

Dr. Barnett is supported by a Clinical Research Fellowship Award by the American Academy of Neurology and the American Brain Foundation.

Word Count:

Abstract: 163
Manuscript: 2128
References: 20
Tables: 2
Figures: 2

Carolina Barnett\textsuperscript{1} conducted the statistical analysis.
Performance of Individual Items of the Quantitative Myasthenia Gravis Score

Abstract

We evaluated the quantitative myasthenia gravis score to determine which items better discriminate among patients and are most responsive to treatment. Current scoring ranges were compared to updated normative data when appropriate. Data from a previous study of intravenous immunoglobulin versus placebo in myasthenia gravis were analyzed. Each item of the quantitative myasthenia gravis score was analyzed for distribution, responsiveness and inter-item correlations. Items on swallowing, speaking and vital capacity had marked floor effects as did grip strength based on current normative data. The timed measures were responsive to treatment but diplopia, vital capacity and grip strength were not. Ocular items were not correlated with bulbar and generalized items (p>0.05), suggesting multi-dimensionality. Contralateral arm and leg tests were highly correlated (r>0.9, p<0.001), indicating redundancy. In summary, we found that the bulbar items and grip strength did not perform well. Timed items from the quantitative myasthenia gravis score, representing fatigability, are responsive to change and we propose new ranges based on updated normative data.

Key Words: Myasthenia Gravis, Outcome Assessment (health care), responsiveness, QMGS
Introduction:

The Quantitative Myasthenia Gravis Score (QMGS) is a commonly used outcome measure in myasthenia gravis (MG). It was first developed by Besinger in 1983 as an outcome measure for a clinical trial[1]. The scale was later modified by Tindal, and was used in a trial of cyclosporine in myasthenia gravis, where treated patients improved on average by 3.5 points compared to the placebo group[2,3]. Barohn in 1998 further modified the scale and performed an inter-observer reliability study[4]. The resulting scale has 13 items, each with a possible score from 0 to 3, and a maximum possible of 39 points, where a higher score indicates more severe disease. The QMGS has been widely used in different trials in myasthenia, and based on the cyclosporine trial, a change in >3.5 units has been used as a marker of meaningful clinical change. More recently, a study of construct validity demonstrated good correlations between the QMGS and different clinical and electrodiagnostic markers[5].

Despite its acceptance as an outcome measure for clinical trials,[6,7] several limitations of the QMGS have been identified. It is time-consuming and requires instrumentation (e.g dynamometer, spirometer). In recent trials of mycophenolate for myasthenia[8,9], some of the items of the QMGS were not as sensitive to change compared to items from other measures, such as the Manual Muscle Test (MMT) and an activities of daily life questionnaire (MG-ADL)[10]. Despite the QMGS having many timed items (continuous data) that are aimed at measuring endurance and hence fatigability, the authors analyzed the items only in terms of the ordinal scores[10]. This is in contrast to other work where the timed items, when measured in seconds as continuous
variables, have been previously shown to be sensitive to change[11]. It is possible that the ranges used by the QMGS, which were determined arbitrarily, don’t accurately reflect clinical change but the categories may also be too coarse to detect change. The aim of this study was to determine which items of the QMGS better discriminate among patients, are more responsive to treatment and whether current scoring ranges are appropriate.

Methods:
Data from a previous study of intravenous immunoglobulin (IVIG) compared to placebo[6] in patients with worsening MG were used. The details of that study have been previously published but in brief, patients were >18 year-old, had a confirmed diagnosis of MG and had worsening symptoms requiring treatment, determined by their treating physician. Patients were randomized to receive IVIG (2g/Kg) or an equivalent volume of saline infusion. The QMGS was the primary outcome measure, and it was assessed at baseline and at day 14 after treatment. For the present study, we analyzed the distributions at baseline of each item using the ordinal scores but also the continuous data when available. For grip strength new normative data have been recently published[12], so the scores at baseline were re-categorized accordingly. For continuous variables, normality was assessed by histograms and normality plots. Inter-item correlations were analyzed by Spearman’s rho, since the distributions were not normal. A correlation coefficient <0.4 was considered low, between 0.4 and 0.8 moderate and ≥0.8 high. Responsiveness was assessed for the total score and for each item (continuous data) by calculating between-groups effect size (ES) with the following formula: [(X change IVIG –
X change Placebo)/ SD of pooled baseline]. The standardized response mean (SRM) was also calculated separately for the IVIG and placebo groups using paired data as follows: 

\[
\frac{\text{Mean score baseline} - \text{Mean score d14}}{\text{SD of the difference}}
\]

A ratio of the SRM for the IVG group compared to placebo was calculated, with ratios >1 indicating change in the treatment groups compared to placebo. For the items without continuous data and only categorical scores, Wilcoxon signed-rank tests were performed to determine change in the median scores after treatment, for each group. Statistical analyses were done with JMP statistical software version 9.0 (SAS Institute Inc., Cary, NC, USA) and R-statistical software version 2.14 (R foundation for statistical computing, Vienna, Austria)[13].

**Results:**

51 patients participated in the original trial, but in one patient the data on the individual items of the QMGS were not available, so this patient was excluded from the analyses. Of 50 patients included, 24 (48\%) received IVIG and 26 (52\%) placebo. The mean age was 56.7 ± 17 years and 20 (40\%) were female. The mean QMGS at baseline was 12.2 ± 4.9 in the IVIG group and 12.4 ± 5.5 in the placebo. There was no difference in the individual items between both groups at baseline. When analyzing the distributions of the scores for each item, we found a marked floor effect (high proportion of patients with a score of 0) for the items on swallowing (78\%), speech (72\%) and vital capacity (66\%) (Figure 1). Grip strength had 20\% and 35\% (left and right) patients scoring 0 using the original QMGS ranges, but using current normative data this increased to 80\% (Figure 2).
In analyzing the inter-item correlations, despite being timed data, the speech item had a binomial distribution with almost all participants developing dysarthria immediately or not at all. Overall the inter-item correlations suggest multi-dimensionality and redundancy in some cases. The items on arms and legs were highly correlated between sides (e.g. right and left arm, r >0.9, p<0.001). The ocular items showed moderate inter-item correlations (r=0.6 p<0.001), but were not correlated with the generalized items or vital capacity (r range:-0.2 to 0.2 p>0.05). The generalized items for head, legs, arms, showed moderate inter-item correlations (r range: 0.6-0.8 p<0.05), and the grip strength and forced vital capacity items were either not correlated or showed poor correlations (r<0.4). Table 1 shows the correlation matrix for the items with continuous data.

When comparing the QMGS total score at baseline and at day 14, the mean change was 2.54 in the treated group and 0.90 in the placebo (p>0.05), with a SRM of 0.52 and 0.15 respectively and a between-groups effect size of 0.34. The items on arm and leg endurance, timed neck flexion and time-to-ptosis had a higher SRM compared to placebo and positive values for between-groups ES. The item on diplopia had a similar SRM in the placebo and treatment groups, with a negative between-groups ES (-0.06), indicating no difference between groups. Table 2 shows the SRM and ES for the complete scale and for each item. For the items on swallowing, facial strength and speech, the Wilcoxon signed-rank test was not significant when comparing the median scores of each treatment group before and after treatment.

**Discussion:**

In this study, we found that the several items of the QMGS, namely the items on swallowing, speaking, vital capacity and grip strength did not have good discriminative
properties between patients, as they have a marked floor effect. The item on grip strength is quite interesting, since when using the scoring ranges of the original QMGS the scores were more evenly distributed, but this dramatically changed when re-categorizing patients with new normative data[12]. This makes clinical sense, since patients with myasthenia usually have a predominance of proximal over distal weakness, so grip strength does not reflect the typical manifestations of MG. The original scoring ranges of grip strength from the QMGS were determined over 20 years ago. The reduction of manual labour through time may explain why normal values of grip strength are lower now than what they were previously. As expected, all of the items with a high floor effect demonstrated poor sensitivity to change, so they are not suitable for measuring change in myasthenic patients. The item on facial strength although evenly distributed at baseline, did not significantly change after treatment in either group. This might be due to the clinical difficulty in differentiating mild, moderate or severe weakness of the orbicularis oculi, and/or the somewhat vague wording of this item in the QMGS.

This study also demonstrates that the timed items for ptosis, arms, legs and head endurance of the QMGS are sensitive to change, when analyzed as continuous data. These findings are in keeping with the findings of Lashley et al.[11], where they studied the clinical response to ephedrine in a group of patients with congenital myasthenia, and showed that endurance of the arms, legs and head significantly improved after treatment. These timed items are ultimately measuring endurance, which is a way to measure fatigability. Fatigability, which is the objective decline in strength after engaging a muscular group in actions normally performed[14], is characteristic of myasthenia, and,
as such, is an important feature to incorporate when assessing disease severity and change in myasthenia patients. The fact that endurance in MG patients improves after treatment supports this approach. Furthermore, timed tests for endurance have demonstrated good inter and intra-observer (test-retest) reliability[15,16], providing more support to their inclusion when evaluating MG patients. Interestingly, the item on diplopia, which is also a timed test, measuring time-to-diplopia, was insensitive to change in this cohort. Although this finding suggests that time-to-diplopia does not accurately reflect clinical change in these patients, we cannot rule out the possibility that diplopia is not particularly responsive to the specific intervention studied (IVIG). The item for diplopia has been used in other scales (it is part of the Myasthenia Gravis Composite [MGC][17]), and it has also been used with the QMGS item on ptosis as a secondary outcome parameter in a trial for ocular MG (Clinical trials registration: NCT00995722). Our study suggests that this item might not be the most accurate way to measure diplopia in MG patients.

Despite the fact that some items were not sensitive to change, the overall QMGS was able to detect change in the treatment group compared to placebo and it is probable that a more significant effect could be demonstrated if new, more responsive items are incorporated. When analyzing the inter-item correlations with continuous data, we found high correlations between sides for the arms and legs. This finding suggests that measuring both sides is redundant, and only one side is necessary. In practice, both arms could be tested simultaneously, using the scoring of the one with the lowest endurance. This also would allow the exclusion of one arm in case of other pathologies that could affect endurance, such as rotator cuff injuries. Testing both legs simultaneously is
physically challenging and it engages the abdominal muscles which limits its feasibility and validity. Viable alternatives might be testing of the dominant leg only or exclusion of a limb where comorbidity (e.g. arthritis) might affect the score. This reduction in number of tests performed has the potential to improve the feasibility and acceptance of the QMGS, since the assessments would be shorter. Furthermore, normative data on healthy individuals aged 18-80 shows that a more appropriate lower limit of normality is 90 seconds for leg lift and 180 seconds for arm abduction[18], and using these normative data as maximum test times would also reduce the overall time necessary to complete the assessments. Applying these upper cutpoints to our study population, we did not find a major floor effect (score 0 changed from 43% to 58% in the arms and from 42% to 46% in the legs). There are also normative studies for the neck flexion test[16,19], where the median time of endurance is about 40 seconds, much less than the 120 seconds of the QMGS. For that test, the minimal clinically important difference has been estimated to be 17 seconds, and this change could also be incorporated when assessing MG patients in the future.

The correlations between the ocular symptoms (ptosis and diplopia) and between the generalized items (arms, legs, neck) were moderate, indicating that they relate to the same domain but are not redundant. However, the ocular items did not correlate to the generalized items, and this suggests multidimensionality, meaning that each domain is independently contributing to the overall construct of disease severity. This should be further studied in other populations, but it makes clinical sense, since in many patients the ocular symptoms do not necessarily go hand in hand with the bulbar and/or generalized
items. Given these considerations, it is possible that each of the dimensions should be scored separately.

In summary, we found that the QMGS timed items for arms, legs and neck endurance, as well as for ptosis discriminate amongst patients with MG and are responsive to change. These items measure fatigability, which is part of the framework of disease severity in MG. The bulbar items, as well as the item on diplopia do not perform well, and new approaches to measuring these items need to be considered to more accurately reflect changes in disease severity and response to interventions. A literature search of other available measures might provide potential items to replace those items from the original QMGS. The oculobulbar facial respiratory score[20] is one example of such a measure. Any changes to the QMGS should be studied prospectively in another population for validation purposes.
References:


8. Muscle Study Group. A trial of mycophenolate mofetil with prednisone as initial


Figure 1. Distribution of Scores at Baseline.

The distributions are based on the item scoring using the QMGS ranges. In this study, Swallowing, Speech and vital capacity hand ceiling effect (>60% patients scoring 0); there was no floor effect for any item (defined as >60% patients scoring 3).
The new ranges were obtained from new normative data by Peters M. et al.\textsuperscript{12} The QMGS range distributions shown are for the left hand; the new normative data doesn’t differentiate between hands, but in this case to allow comparisons, the left hand data was also used. With the new data range, grip strength has marked ceiling effect (80% of patients scoring 0).
<table>
<thead>
<tr>
<th></th>
<th>Diplopia</th>
<th>Ptosis</th>
<th>R. Arm</th>
<th>L. Arm</th>
<th>FVC</th>
<th>R. Grip</th>
<th>L. Grip</th>
<th>Head</th>
<th>R. Leg</th>
<th>L. Leg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diplopia</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ptosis</td>
<td>0.62 *</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R. Arm</td>
<td>-0.03</td>
<td>-0.09</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L. Arm</td>
<td>-0.13</td>
<td>-0.13</td>
<td>0.94 *</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FVC</td>
<td>-0.08</td>
<td>0.07</td>
<td>0.34 &quot;</td>
<td>0.39 #</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R. Grip</td>
<td>-0.1</td>
<td>-0.13</td>
<td>0.23</td>
<td>0.28</td>
<td>0.34</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L. Grip</td>
<td>-0.18</td>
<td>-0.04</td>
<td>0.20</td>
<td>0.28</td>
<td>0.30</td>
<td>0.86 *</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head</td>
<td>-0.23</td>
<td>-0.2</td>
<td>0.65 *</td>
<td>0.68 *</td>
<td>0.28</td>
<td>0.28</td>
<td>0.43 #</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R. Leg</td>
<td>-0.02</td>
<td>-0.17</td>
<td>0.65 *</td>
<td>0.67 *</td>
<td>0.21</td>
<td>0.27</td>
<td>0.27</td>
<td>0.56</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>L. Leg</td>
<td>-0.11</td>
<td>-0.22</td>
<td>0.72 *</td>
<td>0.77 *</td>
<td>0.27</td>
<td>0.36</td>
<td>0.38</td>
<td>0.68</td>
<td>0.90 *</td>
<td>1</td>
</tr>
</tbody>
</table>

All correlations were done by Spearman’s rho. * Denotes p value < 0.0001, # denotes p value <0.01 and ψ p value <0.05.
Diplopia, Ptosis, Arms, Head and Legs were measured in seconds. FVC is measured in % of predicted. Grip strength is measured in Kg.
Table 2. Effect Size (ES) and Standardized Response Mean (SRM) for the QMGS and its items (continuous data)

<table>
<thead>
<tr>
<th>Item</th>
<th>SRM IVIG</th>
<th>SRM Placebo</th>
<th>Ratio SRM IVIG/Placebo</th>
<th>Between Groups Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diplopia</td>
<td>0.13</td>
<td>0.21</td>
<td>0.45</td>
<td>- 0.06</td>
</tr>
<tr>
<td>Ptosis *</td>
<td>0.33</td>
<td>0.05</td>
<td>7.02</td>
<td>0.34</td>
</tr>
<tr>
<td>L. Arm *</td>
<td>0.36</td>
<td>- 0.24</td>
<td>1.40</td>
<td>0.22</td>
</tr>
<tr>
<td>R. Arm *</td>
<td>0.28</td>
<td>- 0.20</td>
<td>1.50</td>
<td>0.18</td>
</tr>
<tr>
<td>FVC</td>
<td>- 0.21</td>
<td>- 0.17</td>
<td>1.24</td>
<td>- 0.03</td>
</tr>
<tr>
<td>R. Grip</td>
<td>0.23</td>
<td>0.27</td>
<td>1.04</td>
<td>- 0.07</td>
</tr>
<tr>
<td>L. Grip</td>
<td>0.23</td>
<td>0.22</td>
<td>0.85</td>
<td>0.00</td>
</tr>
<tr>
<td>Head *</td>
<td>0.70</td>
<td>0.24</td>
<td>2.92</td>
<td>0.12</td>
</tr>
<tr>
<td>R. Leg *</td>
<td>0.32</td>
<td>0.08</td>
<td>4.00</td>
<td>0.16</td>
</tr>
<tr>
<td>L. Leg *</td>
<td>0.26</td>
<td>0.13</td>
<td>2.00</td>
<td>0.23</td>
</tr>
<tr>
<td>Total QMGS #</td>
<td>0.52</td>
<td>0.15</td>
<td>3.47</td>
<td>0.34</td>
</tr>
</tbody>
</table>

The SRM was calculated with paired data using the following formula: 
\[
\frac{(\text{Mean score baseline} - \text{Mean score d14})}{\text{SD of the difference}}
\]. Between groups ES was calculated with this formula: 
\[
\frac{(X \text{ change IVIG} - X \text{ change Placebo})}{\text{SD of pooled baseline}}
\].

* Denotes SRM or ES favouring IVIG, and in this case represents the items most sensitive to change.

# Denotes that both by SRM and ES methods, the IVIG group changed more after treatment than the IVIG group. This indicates that the IVIG as a whole is sensitive to change, despite some items that are not.