Published in final edited form as:

*J Neurol Sci*. 2020 December 15; 419: 117205. doi:10.1016/j.jns.2020.117205.

# It's tricky: Rating alleviating maneuvers in cervical dystonia

**Elizabeth Cisneros**[a], **Glenn T. Stebbins**[b], **Qiyu Chen**[a], **Jeanne P. Vu**[a], **Casey N. Benadof**[a], **Zheng Zhang**[a], **Richard L. Barbano**[c], **Susan H. Fox**[d,e], **Christopher G. Goetz**[b], **Joseph Jankovic**[f], **Hyder A. Jinnah**[g], **Joel S. Perlmutter**[h,i], **Charles H. Adler**[j], **Stewart A. Factor**[k], **Stephen G. Reich**[l], **Ramon Rodriguez**[m], **Lawrence L. Severt**[n], **Natividad P. Stover**[o], **Brian D. Berman**[p], **Cynthia L. Comella**[b], **David A. Peterson**[a,q,*]

[a]Institute for Neural Computation, University of California, San Diego, 9500 Gilman Dr, La Jolla, CA 92093

[b]Department of Neurological Sciences, Rush University Medical Center, 1620 W Harrison St, Chicago, IL 60612

[c]Department of Neurology, University of Rochester, 500 Joseph C. Wilson Blvd, Rochester, NY 14627

[d]Movement Disorder Clinic, Toronto Western Hospital, 399 Bathurst Street, Toronto, ON, M5T 2S8, Canada

[e]Medical Sciences Building, 1 King's College Cir, Toronto, ON M5S 1A8, Canada

[f]Department of Neurology, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030

[g]Departments of Neurology and Human Genetics, Emory University, 1365 Clifton Rd building b suite 2200, Atlanta, GA 30322

[h]Department of Neurology Washington University School of Medicine, 660 S Euclid Ave, St. Louis, MO 63110

[i]Departments of Radiology, Neuroscience, Physical Therapy, and Occupational Therapy, Washington University School of Medicine, 660 S Euclid Ave, St. Louis, MO 63110

[j]Department of Neurology, Mayo Clinic College of Medicine, 200 1st St SW, Rochester, MN 55905

[k]Department of Neurology, Emory University School of Medicine, 201 Dowman Dr, Atlanta, GA 30322

[l]Department of Neurology, University of Maryland Medical Centre, 22 S Greene St, Baltimore, MD 21201

[m]UF Department of Neurology, 1149 Newell Dr, Gainesville, FL 32611

[n]Department of Neurology, Beth Israel Medical Center, 529 W 42nd St # 6K, New York, NY 10036

[*]Corresponding author at.: CNL-S, Salk Institute for Biological Studies, 10010 N. Torrey Pines Rd, La Jolla, CA 92037.

[o]Department of Neurology, The University of Alabama, Tuscaloosa, AL 35487

[p]Department of Neurology, Virginia Commonwealth University, 1101 East Marshall Street, PO Box 980599, Richmond VA 23298-0599

[q]CNL-S, Salk Institute for Biological Studies, 10010 N Torrey Pines Rd , La Jolla, CA 92037

## Abstract

**Objectives—**To investigate hypothesized sources of error when quantifying the effect of the sensory trick in cervical dystonia (CD) with the Toronto Western Spasmodic Torticollis Rating Scale (TWSTRS-2), test strategies to mitigate them, and provide guidance for future research on the sensory trick.

**Methods—**Previous analyses suggested the sensory trick (or "alleviating maneuver", AM) item be removed from the TWSTRS-2 because of its poor clinimetric properties. We hypothesized three sources of clinimetric weakness for rating the AM: 1) whether patients were given sufficient time to demonstrate their AM; 2) whether patients' CD was sufficiently severe for detecting AM efficacy; and 3) whether raters were inadvertently rating the item in reverse of scale instructions. We tested these hypotheses with video recordings and TWSTRS-2 ratings by one "site rater" and a panel of five "video raters" for each of 185 Dystonia Coalition patients with isolated CD.

**Results—**Of 185 patients, 23 (12%) were not permitted sufficient testing time to exhibit an AM, 23 (12%) had baseline CD too mild to allow confident rating of AM effect, and 1 site- and 1 video-rater each rated the AM item with a reverse scoring convention. When these confounds were eliminated in step-wise fashion, the item's clinimetric properties improved.

**Conclusions—**The AM's efficacy can contribute to measuring CD motor severity by addressing identified sources of error during its assessment and rating. Given the AM's sensitive diagnostic and potential pathophysiologic significance, we also provide guidance on modifications to how AMs can be assessed in future CD research.

### Keywords

sensory trick; sensorimotor integration; cervical dystonia; spasmodic torticollis; clinimetrics

## 1. Introduction

A characteristic feature of cervical dystonia (CD) is the "alleviating maneuver" (AM), also referred to as a "sensory trick" or "geste antagoniste" (1–5). The AM transiently reduces the severity of CD motor symptoms and is used by up to 90% of patients (6). It is also a potential clue into the pathophysiology of dystonias, and there is a recognized need to properly measure the AM (7). The AM's efficacy is usually quantified with an item for "Effect of sensory trick" on the motor severity subscale of the Toronto Western Spasmodic Torticollis Rating Scale (TWSTRS (8, 9), TWSTRS-2 (10)). However the item has poor clinimetric properties (11). We hypothesized that this is because of three specific weaknesses. First, patients may not be given enough time to completely demonstrate an AM because there is a mismatch in this regard between the examination protocol and the TWSTRS-2 scoring instructions. Second, it is unclear how to rate the AM's effect for

patients with only slight motor severity because an AM's efficacy is intrinsically scored relative to the patient's motor features at baseline. Third, raters may inadvertently score the AM item in reverse because the TWSTRS-2 scoring anchors are in a direction opposite that of all other items on the scale. With a cohort of 185 CD patients previously evaluated with the TWSTRS-2, we provide evidence supporting these weaknesses, demonstrate how mitigating them improves the item's clinimetric properties, and recommend options for how AM efficacy should be quantified in future CD research.

## 2. Methods

We analyzed clinical and video-based data collected from 208 patients with isolated adult-onset CD enrolled across 10 sites in a rating scale study by the Dystonia Coalition (https://clinicaltrials.gov/ct2/show/NCT01373424). All patients provided informed written consent prior to their inclusion in the study. The protocols for original data collection and subsequent analyses were approved by the Human Research Protection Offices at the Washington University School of Medicine (WUSM), all the local sites, Rush University Medical Center (RUMC), and the University of California, San Diego (UCSD; protocol 111255X). Prior DBS surgery was an exclusion criterion, and patients receiving botulinum neurotoxin (BoNT) were assessed at least three months after their last injection. All patients were video-recorded during a standard examination protocol. Experienced movement disorders neurologists evaluated each patient using the revised Toronto Western Spasmodic Torticollis Rating Scale (TWSTRS-2)(10). The effectiveness of an AM was assessed during a step in the video examination protocol in which patients were seated in a chair without head support, feet resting on the floor, and instructed to demonstrate their most effective AM or to follow three instructions: try touching right cheek, left cheek or back of the head (Table 1). All videographers and rating neurologists received instruction on the examination protocol and the TWSTRS-2 rating scale in both written form and webinar-based training.

All patients were rated live by one rater at their site ("site raters"). Video recordings were screened by two parties blinded to patients' TWSTRS-2 ratings (CB and EC) to eliminate cases in which the videographer did not prompt for an AM as well as cases in which the patient was prompted for an AM but did not try an AM even when prompted. Patients were also excluded if any of the TWSTRS-2 Motor ratings were omitted. All remaining videos were screened for a) "sufficient duration" and b) "sufficient severity". "Sufficient duration" was defined as whether or not the videographer allowed the patient a minimum period of time to demonstrate any individual AM. Although the TWSTRS-2 scoring instructions specified a minimum of 5 seconds for the AM, we found that this was almost never employed, perhaps because the examination protocol instructions did not indicate how long the patient should be asked to demonstrate the trick. As a practical alternative, we chose 2 seconds as a minimum duration adequate to interpret the effect. This was assessed by an independent reviewer (EC) blinded to the patients' TWSTRS-2 ratings. "Sufficient severity" was defined as whether or not the patient was rated greater than one ("slight") on at least one of the predominant postural axes and/or head tremor during baseline conditions by a movement disorders neurologist (CC).

To evaluate inter-rater agreement, out of the original 208 patients, a subset of 80 patients were also selected for TWSTRS-2 rating by a panel of 11 “video raters” (CA, RB, SF, SF, CG, JJ, JP, SR, RR, LS, NS). Every patient was rated by five video raters. The subset of 80 patients was selected to ensure a burden of fewer than 40 patients per rater while ensuring representation from the strata of overall CD motor severity based on global scorings (CC). Raters were assigned patients randomly, with the exception that they were not assigned patients from their own site. All raters assigned to the same patient saw the same video recording of that patient.

For both the site- and video-ratings, we hypothesized that some raters may be using reverse scoring when rating the AM because of the AM item’s counterintuitive rating scale values. As shown in Table 1, for most items on the TWSTRS-2 motor subscale, a higher score corresponds to a more severe motor abnormality, i.e. 0 = none, 1 = slight, 2 = mild, 3 = moderate, and 4 = severe. In contrast, a *more* effective AM is considered to correspond to a *less* severe CD, and therefore assigned a *lower* score, i.e. 0 = complete improvement of CD, 1 = moderate improvement, 2 = mild improvement, 3 = minimal improvement, and 4 = no improvement. As a result, despite the written anchor descriptors, raters may inadvertently use “reverse scoring”, rating the AM such that a more effective AM is scored higher and a less effective AM is scored lower. To identify reverse scoring, we used item response theory (IRT) and generated Test Information Functions using the *mirt* program implemented in R (12). The Test Information Functions display the relationship between an item from the scale (the AM item in this instance) and the latent trait measure of the scale (CD motor severity). There are both conceptual and empirical grounds to suggest that if a person has a high degree of change of CD severity with an AM, this indicates less severe CD, and vice-versa. Conceptually, behind the entire score is the issue of clinical impact from overall CD severity, and if a patient can do something to change the CD intensity, they necessarily have more empowerment over their neurological disability. Empirically, this has also been shown in a study of sensory perceptual discrimination in CD, in which patients with complete AM efficacy had lower overall severity than patients with incomplete AM efficacy (13). Therefore in terms of the latent overall severity of CD, a high AM response should logically be associated with a less severe index of CD. Thus, if the scoring was in the correct direction, the Test Information Functions would normally exhibit a positively sloped sigmodal curve. If the scoring was in the reverse direction, the plot would exhibit a negatively sloped sigmodal curve.

For both the site- and video-ratings, we analyzed the AM item’s clinimetric properties at each of three successive stages of patient cohort selection: 1) the original set of patients, 2) those with sufficient duration and severity, and 3) after having omitted ratings that reflect reverse scoring. For the site-ratings only, in order to use IRT, there was an additional intermediate stage immediately prior to the reverse scoring tests in which we retained patients from only those sites contributing at least 10 patients. To assess the effect of the three successive stages of selection, we examined two measures of internal consistency: the item-to-total correlation (ITC) using the total TWSTRS-2 motor score and change in Cronbach’s alpha if the AM item was removed. ITC values ≥ 0.40 have typically been viewed as adequate (14). Changes in Cronbach’s alpha were identified as either decreased (indicating an improvement in internal consistency), no change, or increased (indicating a

decrease in internal consistency). Additionally, we examined the IRT discrimination value, an indicator of the sensitivity of the item to discriminate high versus low overall CD severity. Discrimination values > 1.00 are considered adequate (15).

For the video-ratings, we also evaluated inter-rater reliability. Because the design was not fully-crossed (not all video raters rated all patients) and because removing reversescoring raters produces a heterogeneous number of raters per patient (5 or fewer), we used a custom measure of inter-rater agreement based on mean absolute difference (MAD, implemented in Matlab; see pseudocode in Appendix A). In brief, MAD aggregates absolute differences among raters across patients, such that a lower MAD corresponds to greater inter-rater agreement. In all statistical tests we used an alpha level of 0.05 to determine significance.

Data Availability Statement: All data available upon reasonable request to the corresponding author.

## 3. Results

Out of 208 patients, the videographer did not prompt for an AM in 8 patients, and 13 who were prompted indicated that they did not have an AM. Of the remaining 187 patients, two patients were omitted because they were missing ratings for head tremor. Of the remaining 185 patients, the TWSTRS-2 Motor total scores had a mean of 16.6, standard deviation of 5.4, and range of 3-29. Of these patients, 17 (9.2%) were identified as having insufficient duration of AM (< 2 sec), 17 (9.2%) were identified as having insufficient severity, and 6 (3.2%) were identified as having both insufficient duration and insufficient severity (i.e. 23 (12.4%) with insufficient duration and 23 (12.4%) of insufficient severity). As a result, 145 (78.4%) of the “site-rated” patients met criteria for sufficient duration and sufficient severity.

The 23 site-rated patients who were not permitted to demonstrate an AM with sufficient duration were not uniformly distributed across sites (Fig. 1). Among sites contributing more than 10 patients, the percentage of site-rated patients with sufficient duration varied between 77-100%.

Of the original 80 video-rated patients, 68 (85%) patients had sufficient duration. Of the 68 patients with sufficient duration, 60 had sufficient severity. When scoring AM efficacy for CD patients who were allowed to demonstrate an AM with sufficient duration, inter-rater variability was higher for patients with insufficient severity (i.e. very mild) than for patients with sufficient severity (Fig. 2; Mann-Whitney rank sum test, medians = 1.14, 0.55; n = 8, 60; $p < 0.005$ two-tailed).

The item response functions represent the item scaling in relation to the latent trait, i.e. total CD motor severity. For example, if an item demonstrated a positive relationship between its scaling and the latent trait, an accelerating function results (e.g., the solid function lines in Figure 3). If, on the other hand, the scaling has a negative relationship to the latent trait, a decelerating function results (e.g., the dashed function lines in Figure 3). In the case of the AM item, an accelerating function would be expected if the item were scored correctly. We found that, among the 7 sites with enough patients to analyze with item response theory, one

of the sites had a rater that was using reverse scoring. Likewise, among the 11 video raters, one was using reverse scoring.

After excluding patients who were not asked to demonstrate their AM with sufficient duration and who were deemed too mild at baseline to rate an accurate change with their AM within the construct of the rating options in the item, the clinimetrics improved for the site and video ratings, as evidenced by increasing item-to-total correlations (ITCs), decreases in Cronbach's alpha if the AM item is omitted, and higher discrimination values from IRT (Table 2). All of these clinimetric properties were improved further after removing the reverse ratings. The conventional minimum thresholds for ITC (i.e. 0.4) and the IRT discrimination value (i.e. 1.0) were exceeded only for the case of multiple video ratings.

## 4. Discussion

### 4.1 Synopsis

If rated systematically and according to predefined criteria, the AM can contribute to the TWSTRS-2 Motor Severity subscale's assessment of CD severity in patients with at least mild symptoms. This finding improves reliability and relationship of the AM item to the latent model of overall CD severity. Nevertheless, the AM item in its present form and application is problematic, and our careful analysis of where the problems reside provides guidance on how to mitigate them in future research.

### 4.2 Our hypothesis tests and how addressing them improves the AM's clinimetric properties

We investigated three potential sources of error in quantifying the AM's efficacy: 1) whether the AM video recording segment was administered according to instructions with respect to a specified duration of AM testing; 2) whether baseline CD severity impacted the ability to evaluate improvement by AM; and 3) whether raters were scoring AM efficacy according to the printed scale or in reverse. We found evidence supporting each of these. First, 12% of the patients in our cohort were given less than 2 seconds to demonstrate the AM's efficacy despite prior training and TWSTRS-2 instructions that direct raters to provide a longer period. This error was not only evident in sites doing minimal recruitment; some of these patients also came from a few sites with the strongest recruitment. This problem can be addressed with added training to support strict protocol adherence and standardized written protocol scripting with instructions such as "*Now that we have discussed and decided on the sensory trick to test on you, please demonstrate it for five full seconds*." This modification could be applied with no actual change to the body of the scale item. Second, not all CD patients exhibit sufficient motor severity to make the effect of the AM detectable according to the currently written options for rating the item. Among the patients in our cohort who were prompted to demonstrate an AM, 12% had a maximal score of "slight" severity ratings on all of the baseline head posture and tremor items. The wording of the item options (ranging from no change to minimal to mild to moderate to complete improvement) is hard to calibrate when the baseline severity is only slight. Among the patients given sufficient time to demonstrate their AM, the video raters exhibited greater agreement for patients with "sufficient severity" than for patients with "insufficient severity". This observation is

consistent with our hypothesis that the difficulty of rating the AM for slightly affected patients leads to greater heterogeneity in AM ratings among multiple raters for the same patient. To correct this problem, the item would need to be revised to include explicit instructions for mild patients. This would then require subsequent validation testing. Third, some raters erred by rating the magnitude of AM efficacy in the reverse order of the actual scale. This is consistent with our hypothesis that the scoring scheme for the AM item – more effective AM is associated with a less severe CD, and vice-versa – is counterintuitive and can confuse raters thereby inadvertently leading to reverse scoring. Like the first problem, this issue can be most efficiently addressed with an alert in the instructions as the rater is about to complete the rating, such as "*NOTE: Unlike other items on the scale, 0 on this item means* ***complete*** *improvement as opposed to* ***no*** *improvement*". Prior to accounting for these issues, the AM item exhibited poor clinimetric properties for both the site- and video-ratings, with item-to-total correlation and IRT discrimination levels well below acceptable thresholds (11). When these issues are successively mitigated by removing patients and raters, the clinimetric properties are markedly improved. This finding implies that the corresponding modifications to the instructions, application, and calibration of the TWSTRS-2 could allow the construct of the AM to be retained within the scale.

### 4.3 Suggestions for retrospective analyses using the TWSTRS-2

Studies that are retrospectively analyzing the TWSTRS-2 Motor Severity subscale using previously acquired patient video-recorded examinations and assessing the AM item in particular should note that previous guidance was to modify or completely eliminate the AM item from TWSTRS-2 because of its low item-to-total correlation, net negative impact on Cronbach's alpha, and very low discrimination value from item response analysis (11). Based on the current study, we would amend that guidance to suggest that investigators should interpret the AM item with caution or incorporate our methods to strengthen the AM item's clinimetric properties, i.e. detect and remove patients with slight severity in whom the effect of the AM is not readily detectable and remove ratings from raters that are scoring in reverse. If the video recordings are available, one can also screen for whether or not patients were given at least 2 seconds to demonstrate the AM. Alternatively, one could simply use a TWSTRS-2 Motor severity total *without* adding the AM item to the total score.

### 4.4 Options for future studies assessing CD: keep the AM item in the TWSTRS-2?

Based on our results, we suggest that future studies wishing to retain the AM item in the TWSTRS-2 incorporate changes in the design of how the item is both prompted and rated. The first matter is procedural. For both in-person and video-recorded examinations, examiners and videographers should be trained to do three things regarding the AM: 1) prompt patients for an AM, 2) if they do not already have one, suggest trying a common AM, and 3) give them sufficient time to demonstrate whether or not (and to what extent) each AM is effective. The minimum duration should be consistent with the TWSTRS-2 scoring instructions, i.e. 5 seconds. The second matter is with respect to rating. Raters should be forewarned of the potential for reverse rating and reminded to refer to the written TWSTRS-2 when rating the AM. Notably these changes relate only to the exam protocol and TWSTRS-2 instructions and do not modify the AM item in the TWSTRS-2 rating scale

itself. The changes would reduce errors in assessing the AM and improve the TWSTRS-2 clinimetric properties.

One of the more challenging issues is the difficulty of rating the effect of the AM in patients with baseline ratings of only slight CD severity. Although most clinical trials include CD of at least minimal severity (e.g. 10-15 on the TWSTRS-2 total severity) (16–19), efficacious treatments could produce cases with very mild CD. In these cases, if the rater is unable to detect a clear difference during an AM, we would suggest that the AM item should be rated as a 0. This would require only a minor modification to the TWSTRS-2 rating instructions for the AM. An alternative would be to add another rating option to the AM such as “CD of slight severity, unable to detect change” (analogous to the MDS-UPDRS). In analyses, this can then be treated as missing data and the overall TWSTRS-2 score prorated, as has been done before with the MDS-UPDRS (20). However, if the scale is changed, it may require re-validating the TWSTRS-2.

Regardless, it may be particularly useful to retain the AM item in the TWSTRS-2 in order to document improvement from baseline for patients that are BoNT naive as a point of reference prior to injections. Also, should patients report the need for ongoing use of an AM, this would be a sign to alter the injection protocol.

### 4.5 Options for future studies assessing CD: remove the AM item from the TWSTRS-2?

A final option for dealing with any problematic rating scale item would be to completely remove it and, in this case, the AM item would simply be dropped from the TWSTRS-2 Motor Severity subscale. This option has some practical and theoretical grounds. The AM item may not be practical because of the need for specific directions that may require extensive training for examiners, videographers, and raters. Although all site-and video-raters involved in our study were experienced movement disorders neurologists, the need to focus specifically on the examination of the AM was not recognized a *priori*. Further, if patients have never heard of or been encouraged to utilize an AM, their demonstrable improvement may be less honed than the improvement seen in patients with well-practiced AMs. Does this mean that their CD severity is actually worse or that their AM is simply less trained? If learning or specific training impacts the AM, having novices evaluated at baseline, then treated with a study preparation and again evaluated, any change in AM could be due to treatment or to practice effect. Without clear knowledge of the effect of learning and practice on AM, the item even if reliable may not be a valid index of treatment change. There are also theoretical grounds to remove the AM item from the TWSTRS-2. Indeed, one might not even necessarily expect AM efficacy to correlate with CD severity. Although the AM is a supportive feature of CD, it is not traditionally considered a motor abnormality *perse*. Rather, as the term “alleviating maneuver” implies, it (transiently) *reduces* motor abnormalities. In this sense, the AM is comparable to a treatment. Severity scales are often used to measure treatment effects, i.e. how severe is the patient before and after a treatment. Thus, the current design that includes the AM as part of the severity assessment is conflating “treatment” and measurement processes. This intrinsically gives rise to 2 out of the 3 problems addressed in this study, i.e. that it is difficult to rate the AM when the patient is very mild and that even experienced movement disorders neurologists with a declared

interest in dystonia are susceptible to inadvertently using reverse scoring when rating the AM. The latter problem arises because of the inverted scoring scheme for the AM item on the TWSTRS-2 Motor Severity scale, i.e. that a *more* effective AM is associated with a *less* severe CD, and vice-versa (see Table 1). This inverted scoring scheme is a natural consequence of two things: 1) that, as discussed above, the AM item is inherently unique relative to other items on the scale and 2) in order to make the AM item rating consistent with the rest of the TWSTRS-2, the rating is based on how much the AM *improves* the abnormal posture. A TWSTRS-2 with the AM item deleted would retain its clinical utility, because previous work has demonstrated that the TWSTRS-2 with the AM item removed still captures the motor severity of CD and retains good to excellent clinimetric properties (11).

Regardless of whether or not the AM is retained as an item in the TWSTRS-2, there should be further work done to explore ways that could more easily evaluate the effect of the AM because of its clinical and pathophysiological significance. The AM is considered a strong diagnostic sign of CD and CD patients have historically suffered from substantial diagnostic delays (21). Furthermore, AM efficacy is predictive of sleep-related quality of life (22), CD patients with an AM have a higher chance of staying employed (23), and it may be possible by optimizing AMs through external devices - such as the use of specialized pens for writer's cramp - to use the AM itself as a therapeutic intervention (24). The presence and impact of an AM may be a useful phenomenon for investigating the underlying mechanisms of CD (24). Successful AMs are associated with a reduction in EMG activity of dystonic muscle and are implicated as an indication of sensorimotor integration (25). Although patients with CD have compromised sensorimotor integration, as evidenced by higher temporal discrimination thresholds (26), the AM can partially mitigate this deficit. For example, in a study of 32 CD patients, those with an efficacious AM were found to have lower visuotactile discrimination thresholds and shorter disease duration, suggesting that an effective AM may be indicative of early adaptive mechanisms of the basal ganglia that are lost as CD progresses (13).

### 4.6 Limitations

There are some limitations of this study. First, we were unable to assess whether a patient was unaware of an effective AM. This could be addressed by a carefully crafted interview and/or exam protocol that would allow demonstration separate from and prior to the examiner suggesting possible AMs to try and carefully documenting cases where the patient reports having been previously unaware of the AM. Second, the sample size in our cohort decreases with each successive step of patient exclusions in our analyses. Despite this, the clinimetric properties for the AM ratings consistently improved over these stages, a process that would otherwise usually adversely affect those clinimetric properties. Third, heterogeneity across patients in terms of their predominant posture, the specific form of their AM, and how the AM might change over time were all outside the scope of this study but would be interesting to evaluate in future studies. Finally, the distribution of TWSTRS-2 scores were moderately skewed in our cohort, with an underrepresentation of the most severe patients (i.e. score of 4). This could limit how well our results generalize to other, particularly more severe, CD patients. Yet, the Dystonia Coalition's recruitment was at

centers of movement disorders expertise where one might expect to recruit a representative proportion of severe patients.

### 4.7 Conclusions

Our study shows that if the AM is performed for an adequate duration, the severity of CD is sufficient to detect changes, and the clinician rates AM efficacy in accordance with the TWSTRS-2 instructions, the clinimetrics of this item markedly improved for both individual ratings and multiple rater ratings. Future assessments of the AM in CD should consider adopting the following strategies: 1) strengthen training on the AM item both in terms of how it is administered during the exam and how it is rated, 2) modify the AM item to include a new option for patients that are deemed too mild to assess the AM, and/or 3) exclude the AM item from the TWSTRS-2 Motor Severity scale and assess it separately.

## Acknowledgments

We gratefully acknowledge Laura Wright and Matt Hicks (WUSM) for assistance with providing data access.

Funding

This research was conducted via the Dystonia Coalition, which is part of the Rare Diseases Clinical Research Network, an initiative funded by the Office of Rare Diseases Research at the National Center for Advancing Translational Sciences (U54 TR001456) in collaboration with the National Institute of Neurological Disorders and Stroke (U54 NS065701) at the National Institute of Health (NIH). This work was also supported by the Office of the Assistant Secretary of Defense for Health Affairs, through the Peer Reviewed Medical Research Program under Award No. W81XWH-17-1-0393. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the Department of Defense.

## Appendix A

Inter-rater agreement measure based on mean absolute difference (“MAD”)

| | |
|---|---|
| R = # of raters | r = rater index ∈ [1,R] |
| P = # of patients | p = patient index ∈ [1,P] |
| $x_{r,p}$ = score by rater r for patient p | |
| M = # of raters per patient (usually 5, but fewer in cases where a reverse rater is omitted) | |

**A.** ∀ rater $r_i$, i = 1: R,

**1.** Identify $\{p_{r_i}\}$ = set of patients scored by rater $r_i$

**2.** ∀ p in $\{p_{r_i}\}$, compute MAD between rater $r_i$ and other raters on patient $p$:

$$d_{r_i,\,p} = \frac{1}{M-1}\sum r_0 = 1{:}\,M, \mathit{excluding}\ i \left| x_{r_i,\,p} - x_{r_0,\,p} \right|$$

**3.**

$$D_{r_i} = \frac{1}{|N|}\sum\{r_i\}\, d_{r_i,\,p}$$

Where N = $|\{p_r\}|$ (e.g. 30 patients)

**B.**

$$MAD = \frac{1}{R}\sum_{i=1}^{R} D_{r_i}$$

## Abbreviations

| | |
|---|---|
| **CD** | cervical dystonia |
| **AM** | alleviating maneuver |
| **TWSTRS** | Toronto Western Spasmodic Torticollis Rating Scale |
| **TWSTRS-2** | Toronto Western Spasmodic Torticollis Rating Scale (REVISED)) |
| **IRT** | Item Response Theory |
| **ITC** | Item-to-Total Correlation |
| **MAD** | Mean Absolute Difference |
| **MDS-UPDRS** | Movement Disorder Society Unified Parkinson's Disease Rating Scale |
| **EMG** | Electromyography |

## References


1. Müller J, Wissel J, Masuhr F, Ebersbach G, Wenning GK, Poewe W. Clinical characteristics of the geste antagoniste in cervical dystonia. J Neurol. 2001;248(6):478–82. [PubMed: 11499637]
2. Patel N, Hanfelt J, Marsh L, Jankovic J, Coalition mot D. Alleviating manoeuvres (sensory tricks) in cervical dystonia. J Neurol Neurosurg Psychiatry. 2014;85(8):882–4. [PubMed: 24828895]
3. Poisson A, Krack P, Thobois S, Loiraud C, Serra G, Vial C, et al. History of the 'geste antagoniste' sign in cervical dystonia. J Neurol. 2012;259(8): 1580–4. [PubMed: 22234840]
4. Gonzalez-Alegre P Descriptions of cervical dystonia by Sir Charles Bell. Mov Disord. 2010;25(3):257–9. [PubMed: 20131384]
5. Defazio G, Albanese A, Pellicciari R, Scaglione CL, Esposito M, Morgante F, et al. Expert recommendations for diagnosing cervical, oromandibular, and limb dystonia. Neurol Sci. 2019;40(1):89–95. [PubMed: 30269178]
6. Jahanshahi M Factors that ameliorate or aggravate spasmodic torticollis. J Neurol Neurosurg Psychiatry. 2000;68(2):227–9. [PubMed: 10644795]
7. Patel N, Jankovic J, Hallett M. Sensory aspects of movement disorders. Lancet Neurol. 2014;13(1): 100–12. [PubMed: 24331796]
8. Consky E, Basinski A, Belle L. The Toronto Western Spasmodic Torticollis Rating Scale (TWSTRS): assessment of validity and inter-rater reliability. Neurology. 1990;40:445.
9. Albanese A, Sorbo FD, Comella C, Jinnah HA, Mink JW, Post B, et al. Dystonia rating scales: critique and recommendations. Mov Disord. 2013;28(7):874–83. [PubMed: 23893443]
10. Comella CL, Fox SH, Bhatia KP, Perlmutter JS, Jinnah HA, Zurowski M, et al. Development of the Comprehensive Cervical Dystonia Rating Scale: Methodology. Mov Disord Clin Pract. 2015;2(2):135–41. [PubMed: 27088112]

11. Comella CL, Perlmutter JS, Jinnah HA, Waliczek TA, Rosen AR, Galpern WR, et al. Clinimetric testing of the comprehensive cervical dystonia rating scale. Mov Disord. 2016;31 (4):563–9. [PubMed: 26971359]
12. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.
13. Kägi G, Katschnig P, Fiorio M, Tinazzi M, Ruge D, Rothwell J, et al. Sensory tricks in primary cervical dystonia depend on visuotactile temporal discrimination. Mov Disord. 2013;28(3):356–61. [PubMed: 23283764]
14. Goodhue DL. Development and Measurement Validity of a Task-Technology Fit Instrument for User Evaluations of Information Systems. Decision Sciences. 1998;29(1):105–38.
15. Baker FB, Kim S- H. Basics of Item Response Theory Using R. 1 ed: Springer; 2017.
16. Truong D, Duane DD, Jankovic J, Singer C, Seeberger LC, Comella CL, et al. Efficacy and safety of botulinum type A toxin (Dysport) in cervical dystonia: results of the first US randomized, double-blind, placebo-controlled study. Mov Disord. 2005;20(7):783–91. [PubMed: 15736159]
17. Comella CL, Jankovic J, Truong DD, Hanschmann A, Grafe S, Group USXCDS. Efficacy and safety of incobotulinumtoxinA (NT 201, XEOMIN®, botulinum neurotoxin type A, without accessory proteins) in patients with cervical dystonia. J Neurol Sci. 2011 ;308(1 −2):103–9. [PubMed: 21764407]
18. Jankovic J, Truong D, Patel AT, Brashear A, Evatt M, Rubio RG, et al. Injectable DaxibotulinumtoxinA in Cervical Dystonia: A Phase 2 Dose-Escalation Multicenter Study. Mov Disord Clin Pract. 2018;5(3):273–82. [PubMed: 30009213]
19. Lew MF, Adornato BT, Duane DD, Dykstra DD, Factor SA, Massey JM, et al. Botulinum toxin type B: a double-blind, placebo-controlled, safety and efficacy study in cervical dystonia. Neurology. 1997;49(3):701–7. [PubMed: 9305326]
20. Goetz CG, Luo S, Wang L, Tilley BC, LaPelle NR, Stebbins GT. Handling missing values in the MDS-UPDRS. Mov Disord. 2015;30(12):1632–8. [PubMed: 25649812]
21. LaHue SC, Albers K, Goldman S, Lo RY, Gu Z, Leimpeter A, et al. Cervical dystonia incidence and diagnostic delay in a multiethnic population. Mov Disord. 2020;35(3):450–6. [PubMed: 31774238]
22. Benadof CN, Cisneros E, Appelbaum MI, Stebbins GT, Comella CL, Peterson DA. Sensory Tricks Are Associated with Higher Sleep-Related Quality of Life in Cervical Dystonia. Tremor Other Hyperkinet Mov (N Y). 2019;9.
23. Molho ES, Stacy M, Gillard P, Charles D, Adler CH, Jankovic J, et al. Impact of Cervical Dystonia on Work Productivity: An Analysis From a Patient Registry. Mov Disord Clin Pract. 2016;3(2):130–8. [PubMed: 27774495]
24. Ramos VF, Karp BI, Hallett M. Tricks in dystonia: ordering the complexity. J Neurol Neurosurg Psychiatry. 2014;85(9):987–93. [PubMed: 24487380]
25. Schramm A, Reiners K, Naumann M. Complex mechanisms of sensory tricks in cervical dystonia. Mov Disord. 2004;19(4):452–8. [PubMed: 15077244]
26. Avanzino L, Cherif A, Crisafulli O, Carbone F, Zenzeri J, Morasso P, et al. Tactile and proprioceptive dysfunction differentiates cervical dystonia with and without tremor. Neurology. 2020;94(6):e639–e50. [PubMed: 31937622]

## Highlights

- Inter-rater variability higher for patients with insufficient motor severity
- Reverse scoring found among TWSTRS-2 alleviating maneuver item ratings
- Sensory trick clinimetric improved when excluding patients with sources of error

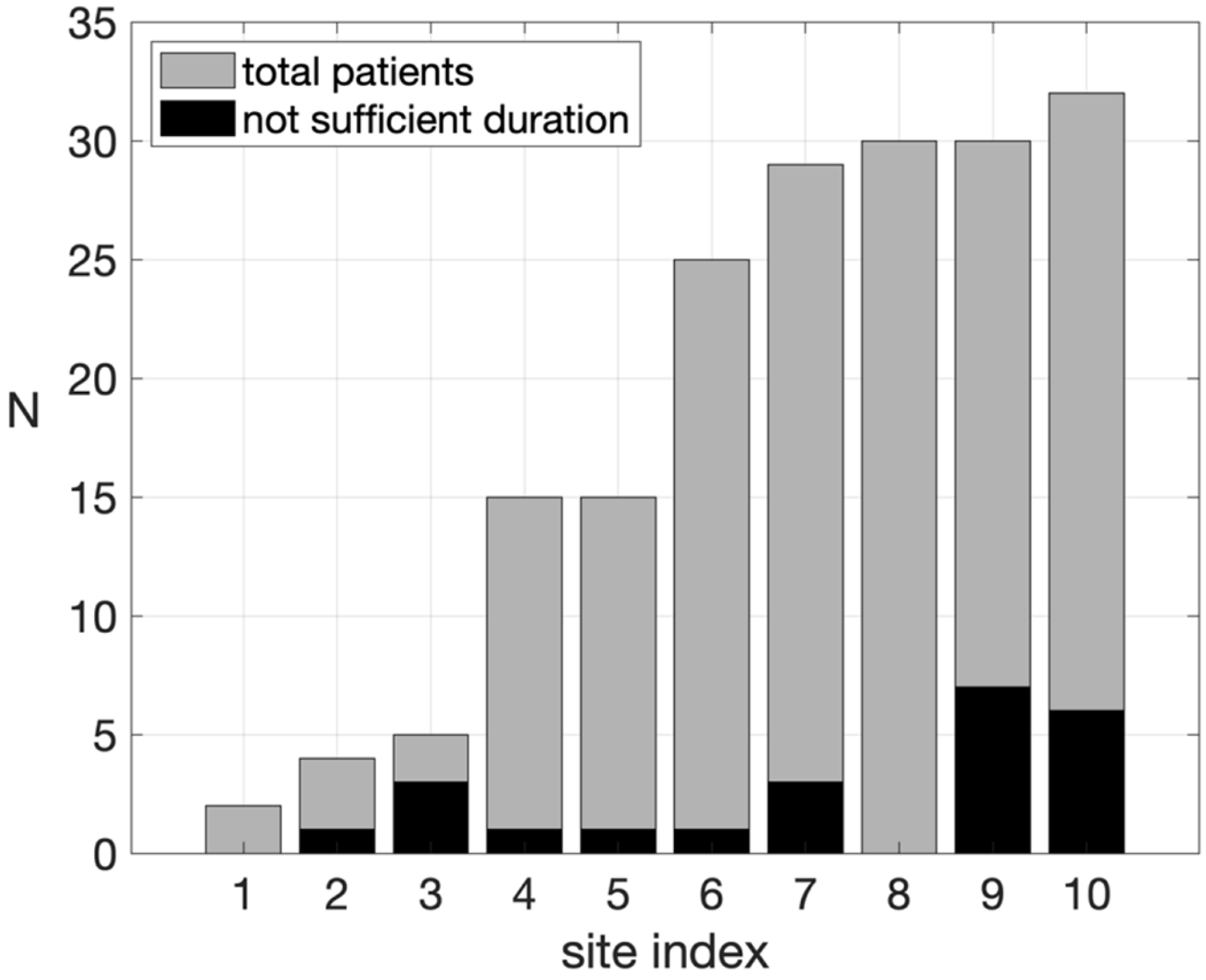


**Fig 1.**

Proportions of patients given sufficient time to demonstrate AM.

Number of patients per site, and number not permitted to demonstrate AM with sufficient duration.

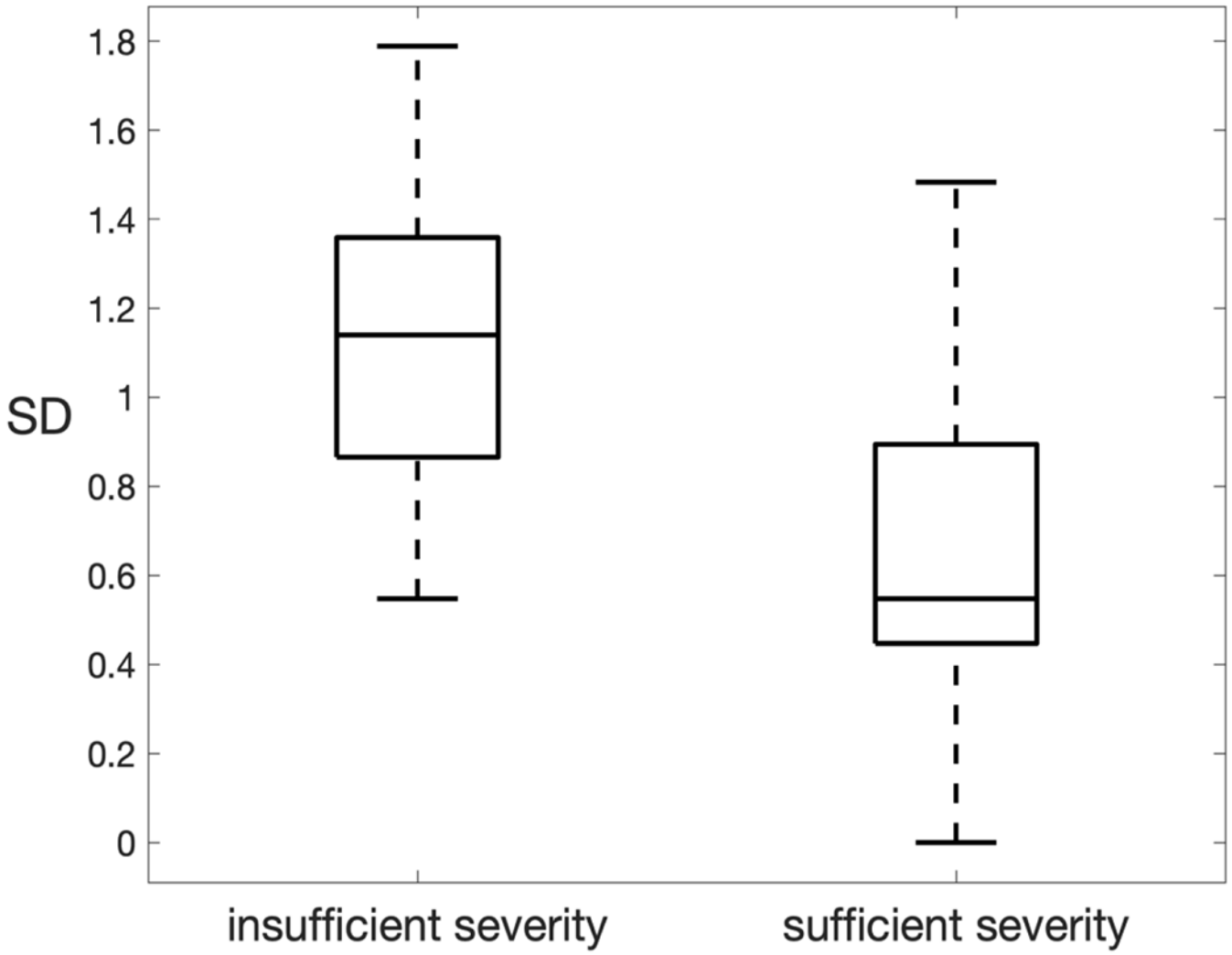


**Fig 2.**

AM rating variability is higher for very mild CD

Inter-rater variability in scoring the AM for insufficiently vs. sufficiently severe patients. (SD = standard deviation among video raters).

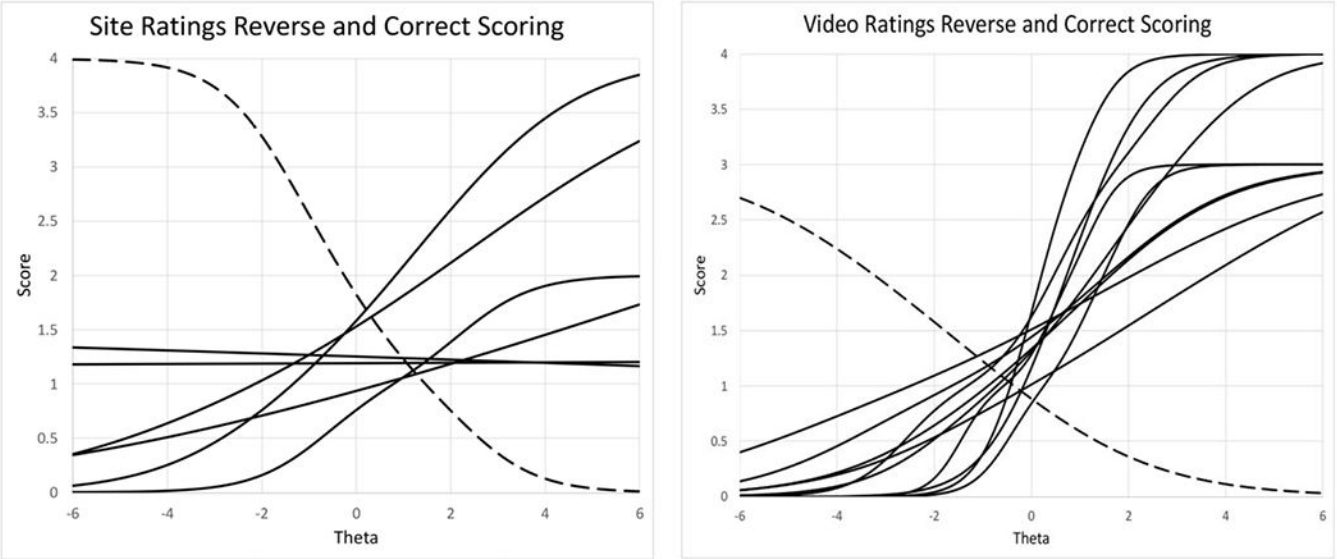


**Figure 3.**

One of each of the site- and video-raters rated AM in reverse of the rating scale.

Test Information Functions for Site Ratings (left) and Video Ratings (right). The x-axis (Theta) represents the severity of the latent trait of CD severity and the y-axis represent the scaling for the AM item. Accelerating curves correspond to the expected direction (Correct Scoring, solid lines) and decelerating curves correspond to the unexpected direction (Reverse Scoring, dashed lines).

**Table 1.**

Examination and rating Instructions for the TWSTRS-2 AM item (and, for comparison, Rotation item)

**Instructions for the videographer during exam:**

*Front view of participant doing most effective sensory trick or a trial of touching right cheek, left cheek and back of head*

**Instructions for the AM item in the TWSTRS-2:**

*A sensory trick is defined as a touch or other movement that influences the severity of the abnormal movements. This item evaluates the degree of improvement when a sensory trick is used. If the patient is unaware of any sensory trick, a trial of touching the cheek, back of the head, or leaning against a wall should be suggested. The improvement in abnormal movements must last at least 5 seconds following application of the trick. If the duration is less than 5 seconds, the score is 4.*

**TWSTRS-2 Motor Severity subscale: the AM item and another example item (Rotation)**

| Sensory Trick | Score | Rotation | Score |
|---|---|---|---|
| Complete improvement of posture by one or more tricks | 0 | None | 0 |
| Moderate improvement of posture by one or more tricks | 1 | Slight | 1 |
| Mild improvement of posture by one or more tricks | 2 | Mild | 2 |
| Minimal improvement of posture by one or more tricks | 3 | Moderate | 3 |
| No improvement of posture by one or more tricks | 4 | Severe | 4 |

**Table 2.**

Clinimetric properties of the AM item for various exclusions

| | Item to Total Correlation | Cronbach's Alpha if Item Omitted | IRT Discrimination | MAD |
|---|---|---|---|---|
| **SITE RATINGS** | | | | |
| Original Sample (n = 208 patients) | 0.011 | Increase | 0.09 | N/A |
| | | | | |
| Sufficient Duration and Severity (n = 145 patients) | 0.153 | Decrease | 0.341 | N/A |
| | | | | |
| Sites with Greater Than 10 Patients (n = 137 patients) | 0.182 | Decrease | 0.409 | N/A |
| | | | | |
| Reverse Scoring Omitted (n = 127 patients) | 0.251 | Decrease | 0.467 | N/A |
| | | | | |
| **VIDEO RATINGS** | | | | |
| Video Ratings with 5 Raters (n = 80 patients, 400 ratings) | 0.231 | No Change | 0.718 | 0.91 |
| | | | | |
| Sufficient Duration and Severity (n = 60 patients, 300 ratings) | 0.405 | Decrease | 1.163 | 0.75 |
| | | | | |
| Reverse Scoring Omitted (n = 60 patients, 273 ratings) | 0.444 | Decrease | 1.29 | 0.68 |

Clinimetric properties of the AM item, before and after excluding patienls with insufficient severity and duration of AM trial and raters using reverse scoring (IRT = item response theory; MAD = mean absolute difference).