Perceptual assessment of dysarthria: Comparison of a general and a detailed assessment protocol

PETTER WANNBERG¹, ELLIKA SCHALLING² & LENA HARTELIUS¹

¹Institute of Neuroscience and Physiology, Department of Health and Rehabilitation, Section of Speech and Language Pathology, Sahlgrenska Academy at the University of Gothenburg, Gothenburg, Sweden, and ²Department of Clinical Science, Intervention and Technology, Division of Speech and Language Pathology, Karolinska Institutet, Stockholm Sweden

Abstract
Objective. The present, preliminary study was designed to investigate whether the results of the use of a detailed assessment protocol ad modum the Mayo Clinic rating of dysarthria and that of a more general assessment protocol, corresponding to ratings of deviations of the different speech production processes, differed primarily in terms of reliability.

Patients and methods. Recordings of text readings of 20 patients with various degrees and types of dysarthria were assessed using both protocols by five clinicians with extensive experience in assessment of neurogenic communication disorders, and results from both assessments were compared.

Results. The general assessment protocol was carried out with higher intra- and inter-rater reliability compared with the detailed assessment protocol. Perceptual deviations were identified in the same domains using both protocols, although only the more detailed protocol could be used to specify particular audible symptoms. Monotony, imprecise consonants, and harsh voice were the most prominent deviations identified with the detailed protocol.

Conclusion. It is concluded that a general assessment protocol is sufficient to identify problem areas reliably and indicate severity of dysarthria but needs to be complemented with a short description of the most prominent audible symptoms and an assessment of intelligibility.

Key words: Dysarthria, intelligibility, perceptual assessment/evaluation, speech, voice

Dysarthria is defined as a collective name for a group of speech disorders resulting from disturbances in muscular control over the speech mechanism due to a lesion in the central or peripheral nervous system (1). Consequently, this group of patients is heterogeneous, and identification of characteristic signs and symptoms has a vital impact on diagnosis and clinical management. Dysarthria is a common consequence of stroke, but can also be the first expression of an underlying neurological disease. The assessment of dysarthria, its presence, type, and severity, can be accomplished in several different ways depending on aim and time available. The methods can vary from having a brief conversation with a patient to find out whether or not a neurogenic speech disorder is present, to carrying out a complete clinical test procedure covering all aspects of speech production, in order to establish the type of dysarthria and possibly aid in the process of diagnosis of neurological disease. In the latter case, a valid and replicable clinical test for dysarthria diagnosis, such as the Frenchay Dysarthria Assessment (FDA-2) is often used (2). Also, any thorough clinical examination of speech is viewed in the broad context of other motor and cognitive signs and the consequences of the speech disorder in terms of decreased intelligibility and restrictions in communicative participation. In addition, blinded as well as not blinded auditory-perceptual evaluations are often used when quantifying and classifying dysarthria for research purposes. The rating scales used in auditory-perceptual analysis are comprehensive and the procedure often time-consuming. The present study describes an effort to investigate if a more general, simplified rating scale can be reliably used to...
confirm the presence of dysarthria and identify problematic aspects of speech production, thus saving time and supplying the clinician/researcher with a perceptually based reliable screening procedure to be used as one additional tool in the assessment of patients in the diagnostic process or to evaluate treatment.

Whether included in a clinical dysarthria test performed in a live meeting with a patient or not, auditory-perceptual evaluation of dysarthria is the cornerstone of the description, quantification, and classification of dysarthria. According to Bunton et al. (3), perceptual evaluation is the ‘gold standard’ for differential diagnosis, judgement of severity, and decisions about management. The prevailing method and classification is the Mayo Clinic rating system (4–6), although the scale and/or the number of dimensions assessed have in some cases been modified. In its original form, the rating system includes 38 perceptual dimensions relating to deviations in respiration, voice, articulation, resonance, prosody, and two overall aspects of speech ( intelligibility and bizarreness). A recorded speaker is rated with regard to each dimension on a seven-point equal-appearing interval scale. The person’s specific constellation of deviant speech dimensions can subsequently be compared with the findings from Darley et al. (4–6) and the speech characteristics identified as a certain type of dysarthria (spastic, flaccid, ataxic, hyperkinetic, hypokinetic, or mixed dysarthria). The perceptual rating scales central to the Mayo Clinic system for assessing different types of dysarthria are perhaps the most comprehensive set of rating dimensions in speech–language pathology.

When conducting perceptual analysis in general, there are mainly three factors that can be varied: 1) the number of listeners, 2) the speech material, and 3) the assessment protocol and procedure, i.e. the perceptual dimensions, the rating scale, and other conditions such as number of presentations. In general, it is preferable to use several listeners rather than one. Listeners are different in their abilities and strategies, and the underlying assumption is that a mean rating score is a more accurate description of the speech deviation compared with the ratings from only one person. Although inter-rater variability is often relatively large, it may decrease with perceptual training (1). The speech material selected for assessment affects which dimensions can be assessed. In the original investigation by Darley et al., (4) the patients were recorded reading a standard paragraph (although it was replaced in some cases with spontaneous speech or sentence repetition). Other types of speech tasks used for assessment of dysarthria are sustained phonation and syllable repetition (alternating motion rates (AMR), i.e. papapapa, tatatata, kakakaka, etc., and sequential motion rates (SMR), i.e. patakapataka etc.). Contextual speech (i.e. reading, conversation, or narration) is considered the most useful task for assessing the integrity of all aspects of speech (1). It has been convincingly argued, e.g. by Weismer (7), that deviations in speech production need to be investigated in contextual speech and not in singular, non-verbal tasks with no meaningful relationship with speech intelligibility. The number of dimensions and scale steps used in a protocol can also be varied. One frequently used modification of the Mayo Clinic system is the one developed by FitzGerald, Murdoch, and Chenery (8). In this revised rating scale the number of perceptual dimensions has been reduced from 38 to 30, and the raters utilize a five-point descriptive equal interval scale (steps 1–5) instead of a seven-point scale to assess the presence and severity of each of the dimensions in samples of text reading. This scale has been used to characterize the speech disorders in patients with e.g. multiple sclerosis, Parkinson’s disease, stroke, and ataxic dysarthria. One of the reasons to reduce the number of scale steps from seven to five is the scale’s correspondence to commonly used terms of severity (normal, mild, moderate, marked, and severe). Other researchers, such as Yorkston, Beukelman, Strand, and Hakel (9) suggest a four-point scale (normal, mild, moderate, and severe) for the assessment of dysarthric signs. The argument for the reduction of the number of scale steps is that the presence of a deviant speech characteristic is more important than its severity (1,9).

The major advantages of perceptual assessment are naturally that the methodology is readily available and has ecological validity in the sense that the perceptual impression of a speaker with dysarthria is closely related to the effectiveness of the communication in terms of intelligibility and communicative adequacy. In addition, Duffy (10) points out that the Mayo Clinic rating system is valuable primarily because 1) it can contribute to localization of neurologic disease (in that different speech symptoms are indicative of the site of lesion within the central or peripheral nervous system), 2) it can and has been used as a point of departure and frame of reference for instrumental research that has increased our knowledge about the characteristics of different dysarthria types and neurologic conditions, 3) it infers underlying deficits such as muscular weakness and incoordination and thus has implications for selection of impairment-focused treatments, and 4) it has provided a vocabulary for a detailed description and communication regarding the perceptual characteristics of dysarthria.

However, several disadvantages have also been reported, particularly concerning low intra- and inter-rater reliability (e.g. 11–13). Reported agree-
ment values vary but have been found to increase with perceptual training and are obviously influenced by the skill and experience of the rater (1). The general finding is that higher agreement values are associated with overall speech dimensions, such as intelligibility and naturalness (14). The more specific the perceptual dimension (such as the ones used by Darley et al. (4)), the lower the agreement. Hence, it seems like the demand for specificity and complexity on the one hand and reliability on the other could counteract each other. The studies mentioned have also shown that the intra-rater as well as the inter-rater variability is different for different perceptual dimensions. Perceptual dimensions that seem to be easy to agree on are: imprecise consonants, excess and equal stress, harsh voice, pitch level, loudness, and fast rate (12,15). Perceptual dimensions that seem to be more difficult to agree on are: irregular articulatory breakdown, distorted vowels, monotonicity in loudness and pitch (because they are inter-related) (12,15).

The Mayo Clinic rating system used for scientific purposes is slightly different from the procedures used for clinical diagnosis. Perceptual analysis for research purposes is analytic and exhaustive, and the listener is focused on assessing the severity of each of the included parameters, which could be described as a bottom-up procedure. Perceptual analysis for clinical diagnosis primarily involves listening for key characteristics with diagnostic value as well as recognizing patterns of perceptual signs, a more top-down procedure. In addition, perceptual analysis for clinical diagnosis is done in the context of a complete dysarthria test (e.g. Frenchay Dysarthria Assessment) (2) where more clinical information is available and each of the speech subsystems (respiration, phonation, oral motor performance, and articulation) is meticulously assessed and where one of the aims is to identify appropriate targets for intervention. Both endeavours are equally important and carried out with great care, but they are different; when discussing the value and merit of perceptual analysis this difference should be kept in mind, and the discussion should relate to the purpose of the analysis and what clinicians actually do in clinical practice.

In summary, auditory-perceptual evaluation of dysarthria ad modum the Mayo Clinic rating system is the primary tool used to describe speech characteristics in research studies and to make decisions about clinical diagnosis. Several aspects of the procedure can be varied, and one aspect that has relevance to reliability and efficiency is the number and scope of the perceptual dimensions used. The present study was designed to investigate whether the results of using a detailed assessment protocol (including the Mayo Clinic rating system dimensions) versus a more general assessment protocol (corresponding to ratings of speech production processes) differ primarily in terms of reliability.

Particular questions asked were:

1. In terms of informativity, do the two different protocols (general versus detailed) generate the same information (albeit at different levels of detail)?
2. Is inter- and intra-rater reliability different between the two protocols?

Materials and methods

Participants and recordings

Twenty consecutive patients with dysarthria, successive visitors to a dysarthria clinic, were included. All patients had given written informed consent regarding the use of recordings for teaching and research purposes without restrictions. Seven of the patients were women and 13 were men, and ages varied between 50 and 84 years (mean 68.8, SD 9.1). All patients were at the time of recording being investigated for various likely or possible neurodegenerative conditions and had no probable or definite neurological diagnosis at the time, except for one patient who was recovering from stroke. They were assessed by a speech-language pathologist at the local university hospital’s neurological clinic. Assessment of dysarthria was made using a Swedish dysarthria test (16), and dysarthria severity was judged as ranging from minimal (two patients), mild (seven patients), mild–moderate (four patients), to moderate (seven patients). No patient was assessed as having severe or profound dysarthria. Types of dysarthria were at this time most often judged to be mixed, with elements of spasticity, hypokinesia, and ataxia. Recordings were made under quiet conditions in the clinic room using digital recording equipment (Zoom H2 or Sony Walkman TCD D-7). The TCD D-7 was used with a head-mounted microphone (AKG C 420) or a table electric condenser microphone (Sony ECM-MS957). The selected speech material for each subject was the reading of a standard text, frequently used by Swedish clinicians for assessments of speech and voice. The text contains 89 words. Six recordings were duplicated for calculation of intra-rater reliability.

Assessment protocols

Two different assessment protocols were created (see Supplementary Appendixes 1 and 2, to be found at online http://informahealthcare.com/doi/abs/10.3109/14015439.2015.1069889). The detailed assessment protocol contained 30 per-
ceptual dimensions, derived from Darley et al. (6) and previously used e.g. in FitzGerald et al. (8) and Hartelius et al. (17,18). The dimensions are classified into five different domains: Voice and breathing (14 dimensions), Resonance (2 dimensions), Articulation (5 dimensions), Prosody (8 dimensions), and Overall assessment of severity of speech deviation (1 dimension). The general assessment protocol included assessment of five parameters, representative of the five domains above. The 30 dimensions of the detailed assessment protocol as well as the five domains of the general assessment protocol were rated using a descriptive interval scale ranging from 0 to 3 where 0 corresponded to (deviances) are non-existent, 1 is a little/seldom, 2 = moderately/sometimes, and 3 is a lot/consistently. The overall assessment of severity of speech deviation ranged from 0 to 3 where 0 = no deviation, 1 = mild, 2 = moderate, and 3 = severe speech deviation. Lists of definitions of all dimensions and domains were created (see Appendixes 1 and 2, available online).

Both assessment protocols were designed as web forms in Google forms, and both protocols were coupled with the 20 + 6 sound files. The duplicated sound files were identical in each of the two assessments (using the detailed and the general assessment protocol), but the randomized order of the sound files differed.

Procedure

Five clinicians, certified speech–language pathologists with extensive (i.e. 10–25 years) experience in assessment of neurogenic communication disorders assessed all sound files. They were informed that the aim of the study was to compare two protocols for assessment of dysarthria but were unaware of the types and severities of the speech disorders and the medical diagnoses associated with the recorded patients.

Written instructions on how to access the forms and sound files via a link and how to fill out the forms online as well as the definitions of dimensions and domains were sent to each of the five listeners. They were also instructed that they were allowed to listen to each sound file as many times as they wanted, within a time-frame of three weeks. Three of the listeners started with the assessment using the detailed protocol (assessing the 20 + 6 recordings) and proceeded to make their assessment using the overall assessment (again assessing the 20 + 6 recordings). Two listeners made their assessments in the reverse order. The sound files only contained the text readings, and both the speakers and the raters were unidentifiable in the data collection process.

The protocols and the sound files were password-protected, and once the assessments had been made the protocols and sound files were removed. The raters were instructed not to download the sound files. All assessment data were automatically transferred to an Excel sheet for transfer to the software for statistical analysis.

Statistical analysis

Statistical analysis was performed using the program R. The level of significance was set to $P<0.05$ for all calculations. In evaluating kappa calculations, the standards of Altman (19) were used. According to Altman, an agreement of $<0.20$ = very weak, 0.21–0.40 = weak, 0.41–0.60 = moderate, 0.61–0.80 = good, and 0.81–1.0 = very good.

Correlations between data generated by the two assessment protocols were calculated using Pearson’s correlation test and are presented visually in graphical plots. These calculations were based on the five raters’ mean rating per domain per sound file for the general assessment protocol and compared with their mean rating of all dimensions per domain for the detailed assessment protocol.

Inter- and intra-rater reliability was calculated using squared kappa. Squared kappa was used because the ordinal data comprised four categories. A $t$ test was used to detect systematical differences in ratings among raters. Marginal homogeneity was calculated as a complement to the $t$ test to detect systematical differences in ratings among raters. The reduplicated sound files were merely used to investigate intra-rater reliability, and only the results of the first presentation of each reduplicated sound file is included in the analysis of results.

Results

Deviant perceptual dimensions and domains

Figure 1 gives an overview of the perceptual characteristics found in the studied group of individuals, as assessed using the two different protocols. Figure 1 shows the mean ratings for each domain and dimension of the general and the detailed protocol, respectively. In the detailed protocol, the dimensions monotony, imprecise consonants, and harshness were the most prominent. Deviations in resonance (hypernasality, mixed nasality) were particularly rare judging by both assessment protocols.

Figure 2 shows a more even distribution of the ratings 0, 1, 2, and 3 for the general protocol compared with the detailed protocol (Figure 3). It is evident from Figure 3 that all raters rated more than
Perceptual assessment of dysarthria

Figure 1. Box plot describing ratings (means and standard deviations across five raters and 20 sound files) for each of the five domains in the general assessment protocol and for each of the 30 dimensions in the detailed assessment protocol. The results of the general assessment protocol’s five domains are on the left hand side of the figure, and the results of the detailed protocol’s 30 dimensions are on the right hand side. The different shades of grey show how the dimensions in the detailed assessment protocol correspond to the domains of the general assessment protocol.

50% of the dimensions as non-deviant (a rating of 0) using the detailed protocol.

Comparison of protocols

Figure 4 shows the raters’ mean ratings for each sound file comparing the general and the detailed assessment protocols, divided into the domains Overall assessment of severity of speech deviation, Articulation, Resonance, Prosody, and Voice and breathing. The values plotted from the detailed assessment protocol are means across the five raters and 30 dimensions. The values plotted from the general assessment protocol are means calculated across five raters and five domains. Every data point represents a particular sound file and a specific domain. Every sound file is represented five times, one time per domain. Values for the detailed assessment protocol are plotted against the y-axis, and values for the general assessment protocol are plotted against the x-axis. Pearson’s $r = 0.64$, $P < 0.001$.

Figure 3. Distribution (percentage) of ratings for each of the five raters when using the detailed assessment protocol.

Figure 4. Comparison of mean ratings for each domain in the general assessment protocol and mean ratings for each dimension in the detailed assessment protocol. All sound files are compared for each domain between the two different assessment protocols, which yields 100 comparisons. Values for the detailed assessment protocol are plotted against the y-axis, and values for the general assessment protocol are plotted against the x-axis. Pearson’s $r = 0.64$, $P < 0.001$. 

Figure 2. Distribution (percentage) of ratings for each of the five raters when using the general assessment protocol.
It is evident from Figure 4 that the ratings were lower when using the detailed assessment protocol than when using the general assessment protocol, which results in a flat slope. This tendency is also visible in a comparison between Figure 2 and Figure 3. However, the overall Severity of speech deviation domain shows a steeper slope and has a similar rating in the detailed and general assessment protocols. The overall Severity of speech deviation domain was also only rated once per sound file and rated in both the detailed and general assessment protocols see Supplementary Appendixes 1–2 to be found at online http://informahealthcare.com/doi/abs/10.3109/14015439.2015.1069889). The strong correlation (Pearson’s $r = 0.64$, $P < 0.001$) between the two different assessment protocols is interpreted as both protocols measuring dysarthric speech and voice difficulties similarly.

**Inter- and intra-rater reliability**

Inter-rater reliability was calculated using squared kappa. All pairwise kappa values were higher for the general assessment protocol than for the detailed assessment protocol (Table I). The mean across the pairwise kappa values for inter-rater reliability was 0.67 for the general assessment protocol and 0.56 for the detailed protocol. Consequently, the reliability values for the general assessment protocol can be considered *good*, while the corresponding values for the detailed protocol are *moderate*.

All raters showed higher intra-rater reliability when using the general assessment protocol, except rater A, who showed the same kappa value regardless of assessment protocol (Table II). Mean intra-rater reliability for the general assessment protocol can be considered *very good*, while the corresponding value for the detailed protocol is *good*.

To investigate systematic differences, all raters were systematically compared using both paired $t$ test and marginal homogeneity (a symmetry test). Significant differences were detected in 13 of 20 pairwise comparisons using $t$ test and in 14 of 20 comparisons using marginal homogeneity (Table III). One obvious case where the symmetry test showed a significant difference and the $t$ test did not is the comparison between raters A and B using the general assessment protocol. This is also visually evident from Figure 2.

Table IV shows how the five raters tended to rate the sound files when using the two different assessment protocols. Mean ratings are consistently higher when using the general assessment protocol.

**Discussion**

In the present study we have compared a type of assessment of dysarthria that can be done relatively quickly (the general assessment protocol) with one that requires more time (the detailed assessment protocol). In summary, the results indicate that the general assessment identified perceptual deviations in the same domains as the detailed assessment in the studied group of speakers, but that it failed to specify the particular audible symptoms. However, the general assessment protocol was carried out with higher intra- and inter-rater reliability compared with the more detailed assessment protocol.

It was stated in the introduction that perceptual evaluation of dysarthria is the prevailing methodology when it comes to making a correct differential diagnosis, assessing severity of deviation, and making decisions about management. So, how do the two different types of assessments compared here contribute in these processes? In identifying the existence of dysarthria, both procedures are probably equally effective, mainly because this decision is based on degree of speech deviation. Differentiating between different types of dysarthria involves the identification of particular key symptoms and combinations of key symptoms. It is conceivable that with the detailed assessment protocol, which lists all possible salient perceptual features, it becomes easier to note and identify the key symptoms of different dysarthria types (i.e. using a bottom-up procedure), which would support the differential diagnosis process. However, identifying dysarthria type is also about detecting patterns (using the top-down

| Rater | General assessment protocol | Detailed assessment protocol |
|-------|-----------------------------|------------------------------|
| A     | 0.91                        | 0.91                         |
| B     | 0.91                        | 0.76                         |
| C     | 0.73                        | 0.70                         |
| D     | 0.83                        | 0.82                         |
| E     | 0.82                        | 0.69                         |
| Mean  | 0.84                        | 0.77                         |

*Table II. Kappa values for intra-rater reliability using squared kappa. For all comparisons $P < 0.001$.***
procedure and e.g. focus the perception on combinations of perceptual features), and a too detailed protocol might contribute to a fragmentation of assessment and impede the synthesizing process. As argued in Duffy (1), there are different styles used for perceptual analysis that can be used in the process of differential diagnosis. Duffy says: ‘What can be missed by this analytic process, however, is the message conveyed by the constant but temporally varying interactions among all of the individual’s normal and abnormal speech characteristics. This appreciation of gestalt cannot be obtained by a checklist approach alone’ (p. 79). In the judgement of severity and decisions about management, it can also be argued that both procedures are equally successful. Judgements of severity assume a holistic approach, and both procedures are able to identify problem areas. Obviously, in a clinical context, no decisions about management are made without detailed information from a clinical dysarthria test, which also supplies a number of tasks and signs that can be used as outcome variables in different intervention approaches.

A couple of factors besides the time aspect and the higher reliability also favour the use of the general protocol as an additional tool in perceptual analysis of dysarthria: There is a well-established high intercorrelation between different dimensions of voice and speech, i.e. the assessment of one dimension is often influenced by others. This in turn indicates a limited ability of the human perceptual system to hear and discern specific details. Sheard et al. (15) found high correlations between all dimensions included in their study, which were imprecise consonants, excess and equal stress, irregular articulatory breakdowns, distorted vowels, and harshness. Harshness tended to correlate least with the other dimensions. It is well known that paying particular attention to deviant details in speech and voice is not in accordance with the human perceptual system. On the contrary, we tend to overlook deviant auditory features in favour of actual understanding of what is being said. Normalization of what you hear is a natural part of auditory perception (20). However, clinicians and researchers in the area of speech–language pathology are trained to disregard the picture on the jigsaw puzzle and focus on the different pieces. As this seems to be counterintuitive, one speculation is that it might be difficult to do it with high enough reliability. Still another reason to use a general assessment is that different perceptual characteristics frequently appear together and are influenced by each other because of a common underlying neuropathological origin.

The present study shows that the detailed assessment protocol generated a large number of 0-ratings. This can be explained by the fact that the recorded individuals, although the severity of dysarthria differed and the group included persons with moderate–severe dysarthria, did not have deviations within all dimensions. The most affected dimensions were monotony, imprecise consonants, and harshness, which are all common perceptual symptoms of dysarthria.

The selected measure of inter-rater reliability was squared kappa. The kappa values of the present study were considered moderate to good (for the detailed assessment protocol) and good to very good (for the general assessment protocol), well above

Table III. Pairwise comparisons of mean difference between the raters using the general assessment protocol and the detailed assessment protocol. The level of significance for the detected difference was calculated using a t test and marginal homogeneity (symmetry). Significant differences in bold.

| Pair     | General assessment protocol | Detailed assessment protocol |
|----------|----------------------------|-----------------------------|
|          | Difference | P (t test) | P (symmetry) | Difference | P (t test) | P (symmetry) |
| A versus B | -0.07    | 0.373  | 0.003       | 0.08      | 0.021     | <0.001  |
| A versus C | -0.37    | <0.001 | 0.001       | 0.03      | 0.417     | <0.001  |
| A versus D | -0.14    | 0.118  | 0.060       | -0.05     | 0.141     | <0.001  |
| A versus E | -0.39    | <0.001 | <0.001      | -0.12     | 0.002     | <0.001  |
| B versus C | -0.30    | <0.001 | 0.003       | -0.05     | 0.117     | 0.241   |
| B versus D | -0.07    | 0.339  | 0.084       | -0.13     | <0.001    | <0.001  |
| B versus E | -0.32    | <0.001 | <0.001      | -0.20     | <0.001    | <0.001  |
| C versus D | 0.23     | 0.007  | 0.029       | -0.08     | 0.010     | 0.055   |
| C versus E | -0.02    | 0.812  | 0.854       | -0.15     | <0.001    | <0.001  |
| D versus E | -0.25    | 0.002  | 0.010       | -0.07     | 0.045     | 0.198   |

Table IV. The five raters’ mean ratings of all domains and dimensions using the general and the detailed assessment protocol respectively.
acceptable norms (0.5–0.9) (18). In similar research, agreement values seem to centre on 0.80 (percentage agreement or correlation coefficient) +/- one scale value (e.g. 2,3,7). The mean kappa for intra-rater reliability was higher than for inter-rater reliability, which is also to be expected. The higher reliability values for the general assessment protocol can, as argued above, be explained by the intercorrelation between perceptual dimensions. It is easier to be in agreement with yourself and others when assessing deviations in a more general or overall perspective. Fewer possibilities and categories to capture the deviations could also be contributing factors. The problem with low reliability in some of the more detailed perceptual protocols is related to the intercorrelation between perceptual dimensions and should be less prominent with assessment from a more general perspective where the domains are less related to each other. However, another possible explanation for the difference between the two protocols’ intra-rater reliability could be that as the time intervals between ratings of the different sound files were shorter when using the general assessment protocol (because it took less time to do the assessment) it was easier for the raters to remember their previous rating of the same sound file. This is a limitation of the present study and should be taken into account in future research.

Systematic differences between the ratings were explored with paired t tests. In 13 of the 20 cases of pairwise comparisons, the differences were statistically significant. Looking at Figures 2 and 3 and comparing with Table III, it is noticeable that some of the pairwise comparisons that were not significant when measured with a t test showed different symmetries, e.g. the comparison between rater A and rater B using the general assessment protocol. The test of marginal homogeneity showed that 14 of the 20 cases of pairwise comparisons were statistically different regarding symmetry. These differences between raters can partly be explained by individual differences in experience, skill, and internal standards. Another aspect is how the rater relates to the scale used. Some raters might tend to keep ratings within the middle range of the scale, while others tend to use all scale steps and have more evenly distributed ratings, which has been previously reported (21). The latter would be an example of difference in method of rating rather than perceptual skill.

To conclude, the present, preliminary study indicates that a general assessment of dysarthria can be made with sufficiently high reliability to identify problem areas and indicate severity of dysarthria. To be useful in a clinical setting, it needs to be complemented with an additional assessment of ‘most audible symptom’ and an assessment of intelligibility. The detailed assessment is more suitable for research purposes. For use in research, particular attention needs to be paid to inter-rater reliability and perceptual training. In addition, a focus on the characteristics and perceptual learning of the raters is particularly warranted.

**Declaration of interest:** The authors report no conflicts of interest.

**References**

1. Duffy JR. Motor speech disorders: substrates, differential diagnosis, and management. St Louis, MI: Elsevier Mosby; 2013.
2. Enderby P, Palmer R. Frenchay dysarthria assessment. 2nd ed. Austin, TX: Pro-Ed; 2008.
3. Bunton K, Kent RD, Duffy JR, Rosenbek JC, Kent JF. Listener agreement for auditory-perceptual ratings of dysarthria. J Speech Lang Hear Res. 2007;50:1481–95.
4. Darley FD, Aronson AE, Brown JR. Differential diagnostic patterns of dysarthria. J Speech Hear Res. 1969;12: 246–69.
5. Darley FD, Aronson AE, Brown JR. Clusters of deviant speech dimensions in the dysarthrias. J Speech Hear Res. 1969;12:462–96.
6. Darley FD, Aronson AE, Brown JR. Motor speech disorders. Philadelphia: WB Saunders; 1975.
7. Weismer G. Philosophy of research in motor speech disorders. Clin Linguist Phon. 2006;20:315–49.
8. FitzGerald FJ, Murdoch BE, Chenery HJ. Multiple sclerosis: associated speech and language disorders. Australian Journal of Human Communication Disorders. 1987; 15:15–33.
9. Yorkston KM, Beukelman DR, Strand EA, Hakel M. Management of motor speech disorders in children and adults. Austin, TX: Pro-Ed; 2010.
10. Duffy JR. History, current practice, and future trends and goals. In: Weismer G, editor. Motor speech disorders: essays for Ray Kent. San Diego, CA: Plural Publishing; 2007. p. 7–56.
11. Kearns KP, Simmons NN. Interobserver reliability and perceptual ratings: more than meets the ear. J Speech Hear Res. 1988;31:131–36.
12. Zeplin J, Kent RD. Reliability of auditory-perceptual scaling of dysarthria. In: Robin DA, Yorkston KM, Beukelman DR, editors. Disorders of motor speech: assessment, treatment, and clinical characterization. Baltimore, MD: Paul H Brookes; 1996. p. 145–54.
13. Zyski BJ, Weisiger BE. Identification of dysarthria types based on perceptual analysis. J Commun Disord. 1987;20:367–78.
14. Linebaugh CW, Wolfe VE. Relationships between articulation rate, intelligibility and naturalness in spastic and ataxic speakers. In: McNeil MR, Rosenbek JR, Aronson AE, editors. The dysarthrias: physiology, acoustics, perception, management. San Diego, CA: College-Hill Press; 1984. p. 197–205.
15. Sheard C, Adams RD, Davis PJ. Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility. J Speech Hear Res. 1991;34:285–93.
Supplementary material available online

Appendix 1–2.