JSLHR

Research Article

Gauging the Auditory Dimensions of Dysarthric Impairment: Reliability and Construct Validity of the Bogenhausen Dysarthria Scales (BoDyS)

Wolfram Ziegler,[a] Anja Staiger,[a] Theresa Schölderle,[a] and Mathias Vogel[b]

Purpose: Standardized clinical assessment of dysarthria is essential for management and research. We present a new, fully standardized dysarthria assessment, the Bogenhausen Dysarthria Scales (BoDyS). The measurement model of the BoDyS is based on auditory evaluations of connected speech using 9 scales (traits) assessed by 4 elicitation methods. Analyses of the BoDyS' reliability and construct validity were performed to test this model, with the aim of gauging the auditory dimensions of speech impairment in dysarthria.
Method: Interrater agreement was examined in 70 persons with dysarthria. Construct validity was examined in 190 persons with dysarthria using a multitrait-multimethod design with confirmatory factor analysis.
Results: Interrater agreement of < 1 on a 5-point scale was found in 91% of cases across listener pairs and scales. Average reliability was .85. Inspection of the multitrait-multimethod matrix pointed at a high convergent and discriminant validity. Modeling of the BoDyS trait and method factors using confirmatory factor analysis yielded high goodness of fit. Model coefficients confirmed high discriminant and convergent validity and revealed meaningful relationships between scales and methods.
Conclusions: The 9 auditory scales of the BoDyS provide a reliable and valid profile of dysarthric impairment. They permit standardized measurement of clinically relevant dimensions of dysarthric speech.

Clinical assessment of dysarthria is a slowly developing area. The diagnostic approach proposed more than 40 years ago by the Mayo Clinic group around Darley (e.g., Darley, Aronson, & Brown, 1975) is still influential today. Following this approach, a clinician evaluates a speech sample elicited from a patient (i.e., reading of a standard text) by rating each of a list of 38 auditory parameters on a 7-point scale. The test variables refer to signs of impaired respiratory (e.g., audible inspiration), phonatory (e.g., harsh voice), articulatory and resonance (e.g., imprecise consonants, hypernasality), or prosodic functions (e.g., reduced stress), as well as to overarching dimensions such as intelligibility and bizarreness. Due to its comprehensiveness and its groundbreaking role in the understanding and classification of the dysarthrias, the Mayo Clinic rating system has had an enormous impact on dysarthria research (Duffy, 2013; Duffy & Kent, 2001). However, this approach has not played a role as a standardized clinical assessment tool so far, presumably due to its extensive number of test variables and its limited reliability. Bunton, Kent, Duffy, Rosenbek, and Kent (2007), for instance—who analyzed between-listeners agreements across the 38 dimensions of the Mayo Clinic system—found that one third of their rating pairs differed by 2 scale points or more on the 7-point scale used in this system. An average of 17% of the ratings collected in this study even differed by 3 points or more—that is, by at least one half of the scale length. Other reports have documented results that were still more disappointing (e.g., Zyski & Weisiger, 1987). In general, there is considerable skepticism concerning the reliability of auditory-perceptual ratings in the clinical assessment of the dysarthrias (for discussions, see, e.g., Bunton et al., 2007; Chenery, 1998; Duffy & Kent, 2001; Kent, 1996, 2009). One of the major challenges, especially of the Mayo Clinic rating system, is that listeners may have problems distinguishing sharply enough between the many

[a] Clinical Neuropsychology Research Group, Institute of Phonetics and Speech Processing, Ludwig-Maximilians-Universität, Munich, Germany
[b] Clinic for Neurology, Neurophysiology, Neuropsychology, and Stroke Unit, Clinic Bogenhausen, City Hospital Munich, Germany

Correspondence to Wolfram Ziegler: wolfram.ziegler@ekn-muenchen.de

Editor: Julie Liss
Associate Editor: Tanya Eadie

Received August 23, 2016
Revision received December 8, 2016
Accepted January 28, 2017
https://doi.org/10.1044/2017_JSLHR-S-16-0336

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

concurrent dimensions of deviant speech in their evaluation of a speech sample (Duffy & Kent, 2001; Sheard, Adams, & Davis, 1991), which may delimit not only the reliability but also the discriminant validity of this assessment.

Among the attempts to overcome this dilemma, essentially two different alternatives have been proposed. One alternative approach counts on objective measures instead of perceptual ratings, primarily on acoustic parameters. Great efforts have been undertaken over decades to implement acoustic methods as a clinical assessment tool, because these methods contain the promise of being incisively objective and reliable (e.g., Forrest & Weismer, 2009; Kent & Kim, 2003). However, to date there is still no satisfying solution to the problem of how such a tool can be designed in a way that it provides clinically valid and practically relevant data across the broad range of severity levels and symptom patterns seen in a clinic or private practice. For instance, voice onset time measures of plosive articulation (e.g., Özsancak, Auzou, & Hannequin, 2001) are not applicable in patients who are unable to produce a complete vocal-tract closure, and measures of vowel formant space (Lansford & Liss, 2014; Ziegler & von Cramon, 1983) may become invalid in the presence of strong nasal resonance or aberrant voice quality (Derdemezis et al., 2016). Hence, acoustic parameters play an important role as selectively useful diagnostic adjuvants, but they are still far from serving as a broadband tool for standardized clinical testing.

A second approach toward developing standard clinical assessment tools still relies on mainly perceptual analysis methods, but seeks to increase the reliability and user-friendliness of the Mayo Clinic rating system by using a less fine-grained inventory of test variables and a greater diversity of tasks. Most representative of this type of "composite examination instrument" (Kent, 2009) is the Frenchay Dysarthria Assessment–Second Edition (FDA-2; Enderby & Palmer, 2008), which has gained a wide distribution internationally. Similar clinical dysarthria tests or adaptations of the FDA-2 have been published in various languages, such as the Dysartritest in Swedish (Hartelius & Svensson, 1990), the Battery d'Évaluation Clinique de la Dysarthrie in French (Auzou & Rolland-Monnoury, 2006), or the German Frenchay Dysarthrie-Assessment–2 (Enderby & Palmer, 2012). Composite examination instruments of this kind are based on tasks such as food intake (e.g., eating a cookie, drinking a glass of water), respiratory-vocal maneuvers (e.g., sniffing, coughing, throat clearing), intra-oral airstream maneuvers (e.g., smacking, puffing out the cheeks), lingual-labial motor skills (e.g., lateral tongue movements, lip pursing), maximum performance tasks (e.g., rapid sequential syllable repetitions or sustained vowels), and speaking. Of note, speech is assessed by only eight out of 26 tasks of the FDA-2 protocol. Moreover, the FDA-2 includes a diversity of assessment methods, such as visual inspection of the articulators at rest, anamnestic queries, time measurements using a stopwatch, and auditory judgments of speech parameters (Enderby & Palmer, 2008).

The problem with such composite task designs of dysarthria assessment is that they lack a model of their supposedly underlying theoretical constructs and of how these assumed latent factors are related to the alleged overall test construct—that is, dysarthria. Because tasks such as throat clearing, lateral tongue movement, eating a cookie, and vowel prolongation lack any face evidence as indicators of a speech problem, provision of empirical evidence confirming their validity as measures of some dysarthria-related latent factor is a crucial requirement. Moreover, against the background of continuing discussions on the relationship between speech and nonspeech functions of the respiratory, laryngeal, and vocal-tract muscles (Kent, 2015; Weismer, 2006; Ziegler, 2003; Ziegler & Ackermann, 2013), empirical validation of such examinations is imperative for test-theoretical reasons. As an example, Staiger, Schölderle, Brendel, Bötzel, and Ziegler (2016) recently conducted a factor-analytic investigation of the relationship between tasks eliciting speech on the one hand and nonspeech or paraspeech articulator movements on the other, demonstrating that these different task types actually represent different latent traits. It is also known that impairments of speech and nonspeech or paraspeech vocal-tract movements may dissociate (Staiger, Schölderle, Brendel, & Ziegler, 2017). In consequence, inclusion of nonspeech or paraspeech tasks in a standardized assessment battery of dysarthria would need to be grounded on a theoretically motivated and empirically attested measurement model explaining the relationship of these tasks with the construct of a dysarthric impairment of speaking and communicating.

Furthermore, comparison of a person's performance across the different tasks of a multidimensional assessment requires transformation of the raw test scores into standard norms and computation of separate confidence intervals for each test variable (Cohen, Swerdlik, & Sturman, 2013). To our knowledge, these requirements are not fulfilled by any of the dysarthria assessments published so far.

In the face of such obvious standardization problems, common clinical practice relies mostly on nonstandardized collections of speech and nonspeech performance tasks for the assessment of diverse motor functions of the musculature engaged in speaking. Kent (2009), for example, recommended the use of an assessment protocol menu for the compilation of customized examination protocols, by selecting tasks that are deemed suitable for the assessment of a patient's individual speech motor problem. In a similar vein, Duffy (2013, Chapter 3) proposed a general-purpose clinical examination tool that is based on a nonstandardized inventory of auditory-based motor speech examinations, including maximum performance tasks, together with a formalized evaluation protocol. These approaches are based on the idea that the provision of a large toolbox of motor tasks can help therapists in gaining a deeper understanding of their patients' speech impairments. Reliability and ecological validity are hoped to ensue as a side effect of the intensive interaction between clinicians and patients during the course of condition management (Duffy, 2013, p. 78).

There is no doubt that this type of individually tailored approach is clinically valuable, especially when it comes to questions of how a patient should be treated.

Nonetheless, in other regards such examinations also have serious limitations: To the extent that assessment parameters used in these examination tools are not standardized, they do not permit reliable comparisons across time, across patients, or across the dimensions of a patient's dysarthria profile. Hence, there is still a strong need for reliable and valid assessment instruments and for standard norms in the diagnostic process, not only for the documentation of the pattern and severity of dysarthria and its course in a patient's clinical records but also for use in dysarthria research.

This article reports on a new instrument for the standardized clinical assessment of dysarthria: the Bogenhausen Dysarthria Scales (BoDyS; Nicola, Ziegler, & Vogel, 2004; Ziegler, Schölderle, Staiger, & Vogel, 2015). The BoDyS is a German-based test that involves exclusively speech tasks as its test items and perceptual ratings as its measurement method. The test has already undergone a variety of standardization procedures and has been applied in several investigations of dysarthria (see later). In the current study, we expand more deeply on the measurement model underlying the BoDyS, especially on the questions of whether the nine BoDyS variables represent different constructs, how they are related to each other, and whether they are invariant under different elicitation modalities (construct validity). Because reliability is a necessary precondition of validity, the nine BoDyS scales were also examined for their reliability.

The article is organized as follows: To familiarize readers with the test, the next section outlines the objectives and construction of the BoDyS. The following two sections deal with reliability and validity issues, each with a brief outline of the research question and a short discussion. The final section presents conclusions and implications.

BoDyS

Objectives and Design

The BoDyS is an instrument for the standardized clinical assessment of dysarthria. Comprehensive descriptions of the test have been published in German (Nicola et al., 2004; Ziegler et al., 2015), whereas shorter outlines can be found in research articles by Brendel et al. (2013, 2015), Schölderle, Staiger, Lampe, and Ziegler (2013), and Schölderle, Staiger, Lampe, Strecker, and Ziegler (2016). The major objective of the test is to assess the severity of dysarthric impairment on a multidimensional scale, with the aim of providing a psychometrically sound basis for the planning of therapeutic interventions and the measurement of progression, recovery, or treatment effects. As such, the BoDyS is also considered useful as a research tool for the investigation of speech motor capabilities in different etiologic groups or in longitudinal trials (Brendel et al., 2013, 2015; Schölderle et al., 2016).

Though classification of dysarthria subtypes is not among the objectives of the BoDyS as a standardized measurement tool, the test also comprises, in parallel with the core dysarthria scales, a list of dysarthria features that may be used for purposes of classifying dysarthria subtypes (Schölderle et al., 2013) and designing individualized treatment plans. In this regard, the BoDyS offers similar options as the Mayo Clinic rating system.

Elicitation Methods

The test involves elicitation of 12 samples of connected speech using four different elicitation methods (see Table 1): interview questions ("conversational speech," CONV; three samples), sentence repetition (SENT; three lists of five sentences each), passage reading (READ; three short text passages), and narration of a picture story (PICT; three stories comprising four pictures each). These different elicitation methods are used to cover a range of conditions involving different cognitive and motor demands. Cognitive requirements are known to influence motor speech functioning in healthy individuals (Kleinow & Smith, 2006) and in neurologic populations (e.g., Huber & Darling, 2011). The four types of speech tasks used in the BoDyS protocol vary along at least two dimensions: the requirement of formulating a coherent text (high in CONV and PICT, low in SENT and READ) and utterance length or the option to organize utterances into manageable portions (favorable in CONV and PICT, unfavorable in READ and the longer sentences of the SENT task).

Scales and Features

Scales

The BoDyS variables comprise nine scales representing major dimensions of potential dysarthric impairment: respiration (RSP), voice level (VOL), voice quality (VOQ), voice stability (VOS), articulation (ART), nasal resonance (RSN), articulation rate (TEM), fluency (FLU), and prosodic modulation (MOD). They cover the three motor

Table 1. The four BoDyS speech elicitation methods.

Elicitation method	Description
CONV (Conversational speech)	Three open interview questions that ask about occupation and education, leisure and hobbies, and vacation and holiday destinations
SENT (Sentence repetition)	Three sets of five sentences each, controlled for syllabic length (4–12 syllables) and mode (declarative, interrogative, imperative, subordinate-clause constructions)
READ (Text reading)	Three texts (between 80 and 90 words each) on daily life topics, avoiding complex grammar and foreign words
PICT (Picture story)	Three collections of four cartoon pictures telling a short story with a humorous punch line

Note. Each task type is represented by three tasks, amounting to a total of 12 speech samples.

components involved in speaking—that is, the respiratory, laryngeal, and supralaryngeal systems—as well as three prosodic dimensions that are commonly accepted as being implicated in dysarthria. Table 2 provides a list of the nine BoDyS scales.

For each of the 12 speech samples, each scale is awarded a score between 0 and 4, with 4 denoting complete absence of any impairment and 0 denoting most severe dysfunction. Overall, across the 12 test items an individual is assigned between 0 and 48 points on each scale, yielding a dysarthria profile that indicates severity of impairment across the nine dimensions listed in Table 2.

The BoDyS scales constitute the core of the BoDyS as a psychometrically approved assessment tool. They are conceived of as test variables representing different underlying constructs (latent traits). There is, notably, no one-to-one correspondence between scales on the one hand and circumscribed speech motor systems or motor mechanisms on the other. As an example, the loudness aspect of the VOL scale is associated as much with the respiratory as the laryngeal system. Furthermore, there are various interactions between the motor systems associated with different scales. For instance, velar insufficiency (as measured primarily by the RSN scale) may, as a consequence of excessive loss of air, cause secondary symptoms at the RSP level (frequent inspirations) and/or at the VOQ level (strained-strangled voice due to excessive laryngeal adduction for the prevention of air loss). Also, the factors underlying the three prosodic scales—that is, TEM, FLU, and MOD—notably must necessarily interact with the respiratory, voice, and articulation scales, because the prosodic quality of dysarthric speech is a function of the integrity of the motor subsystems. Hence, the BoDyS scales should be understood as variables representing circumscribed auditory surface phenomena with a high transparency for their underlying motor mechanisms.

Table 2. The nine BoDyS scales and their associated features.

Scale	Features
RSP (respiration)	Frequent inspirations Audibly effortful inspirations Expiration beyond resting level
VOL (voice level)	High pitch Low pitch Loud voice Soft voice
VOQ (voice quality)	Breathy and harsh Strained-strangled and harsh
VOS (voice stability)	Fluctuations of pitch, loudness, and/or quality Voice breaks and voice fading Voice tremor Involuntary vocalizations
ART (articulation)	Reduced Overshooting Wide Narrow Variable
RSN (resonance)	Hypernasality and nasal emission Hyponasality Intermittent hyper- or hyponasality
TEM (articulation rate)	Decreased articulation rate Increased articulation rate
FLU (fluency)	Pauses Iterations
MOD (prosodic modulation)	Reduced prosodic modulation of pitch and/or loudness Excessive prosodic modulation of pitch and/or loudness Conspicuous rhythm and/or stress patterns

Features

Each BoDyS scale is described by a set of two or more dysarthric features, resulting in an overall list of 29 features (see Table 2, right column). Each feature is associated with one of the nine BoDyS scales, thereby contributing to a prescription of the construct represented by the respective scale. The BoDyS features are roughly comparable with the deviant speech dimensions of the Mayo Clinic system (Darley et al., 1975).[1]

Unlike the BoDyS scales, the 29 BoDyS features are not subject to any scaling or standardization processes, but rather are rated only for their presence or absence. Observation of the presence of a feature in a speech sample is registered by checking the respective feature on an evaluation form, in order for the examiner to collect and document information about the quality of the individual's limitations in each dimension. Up to 12 ticks can be placed for each feature, one for each speech sample. The number of ticks indicates the consistency of occurrence of a feature across the whole testing but not its severity. Although such information can be useful in classifying an individual's dysarthria (Schölderle et al., 2013) or determining treatment goals, feature-related BoDyS data are not licensed by psychometric analyses and will not be considered any further in this article.

Test Administration and Evaluation

The test is administered in a quiet room. The individual's utterances are recorded for offline evaluation, preferably on videotape. The three parallel forms are examined blockwise. Within each of the three blocks, the four elicitation modalities are administered in the order given in Table 1. Testing time is approximately 15–30 min.

The 12 speech samples are evaluated sequentially. During and after the playback of a sample, the examiner checks the deviant features on a form and then awards a score for each scale. The individual scores of the 12 samples are added to obtain a sum score for each scale. Moreover, a weighted grand average (profile height) can be calculated across the standardized scores of all scales to obtain a total BoDyS score as a measure of overall dysarthric impairment.

[1] One feature that is not contained in the Mayo Clinic rating system or in other catalogues of dysarthria dimensions is the wide–narrow distinction of the BoDyS ART scale. This feature was included in the BoDyS feature list to permit documentation of two diagnostically and therapeutically relevant symptoms: increased and reduced jaw opening.

In their interpretations of BoDyS profiles, examiners should keep in mind that the scores reflect surface manifestations of an individual's dysarthria rather than direct measures of physiological-level deficits. Any conclusions regarding underlying pathomechanisms require consideration of the potential interactions between speech motor subsystems and of how they may surface as audible symptoms.

Standardization

Test reliability has already been examined in several preliminary studies (Brendel et al., 2013; Ziegler et al., 2015). Furthermore, various aspects of criterion validity have been reported. Among them are findings of high correlations between BoDyS scores and associated acoustic measures (Brendel et al., 2013) and of a potential of the BoDyS to map differences between etiologic groups (Brendel et al., 2015; Schölderle et al., 2013). Moreover, a high predictive value of BoDyS scores has been demonstrated for measures of naturalness and intelligibility (Brendel et al., 2013; Schölderle et al., 2016), self-reports of individuals with dysarthria (Hinterberger et al., 2008), and the attitudes of laypersons toward individuals with dysarthria (Schölderle, Staiger, Heß, & Ziegler, 2017). At present, standard T-norms derived from BoDyS profiles of 210 individuals with dysarthria are available. They permit application of parametric tests for comparisons between and within BoDyS profiles.

Normative BoDyS scores were obtained from a sample of 70 neurologically healthy adults across a broad age range, in order to obtain cutoff scores for the detection of impaired performance on each scale. Protti (2014) prepared an Italian version of the test. A list of all of the cohorts using the BoDyS for psychometric and clinical purposes is included in the Appendix.

The Status of the BoDyS Within a Broader Diagnostic Framework

Within a broader diagnostic framework, the BoDyS has the role of a core module providing standardized numerical scores for a representative set of auditory dimensions of dysarthric impairment. BoDyS profiles contain reliable cues to an individual's core deficits relating to respiration, voice, articulation, and prosody in connected speech. The profiles also allow for comparisons between scales, between speakers, and between follow-up measurements using a statistically sound method.

The BoDyS as a measure of functional impairment does not replace assessment of an individual's intelligibility or other communication-related variables, or of the individual social consequences of the disorder. Evaluation of these parameters requires separate diagnostic tools that are customized for these particular purposes. On this account, the diagnostic framework in which the BoDyS is embedded conforms to the International Classification of Functioning, Disability and Health concept of the World Health Organization (2001), which distinguishes between the consequences of a health condition on function, activity, and participation. The level of intelligibility for a person with dysarthria especially deserves examination by a separate standardized test (e.g., Ziegler & Zierdt, 2008), due not only to its outstanding relevance but also to the specific methodological requirements of intelligibility testing—for example, regarding the control of listener familiarity effects (Liss, Spitzer, Caviness, & Adler, 2002). As mentioned earlier, however, the BoDyS variables are powerful predictors of a number of communication-relevant parameters, demonstrating the content validity of the BoDyS approach (e.g., Schölderle et al., 2016).

Furthermore, administration of the BoDyS will usually not preclude clinicians from performing further, customized examinations to explore the individual resources or specific aspects of speech impairment of a person with dysarthria, especially for purposes of differential diagnosis, treatment planning, or decision making. The standardized BoDyS profile along with the BoDyS feature counts provide information that may be valuable in guiding such additional explanatory diagnostics.

Reliability

Objectives

Reliability is a necessary precondition of construct validity, and high reliability coefficients are required to obtain sufficiently high measurement accuracy—that is, sufficiently narrow confidence intervals for the test scores (Cohen et al., 2013). As described in the introduction, one of the major challenges of auditory evaluation methods in dysarthria assessment is that listeners may have fuzzy concepts of the criteria underlying the variables to be rated. Such auditory misconceptions are considered to contribute to the problem of limited test reliability, especially if the number of dimensions to be distinguished is large and the number of test items is small, as in existing dysarthria tests (Duffy & Kent, 2001; Wannberg, Schalling, & Hartelius, 2016). The BoDyS was designed to mitigate this problem by limiting the number of test variables to nine and by basing each score of the BoDyS profile on 12 items. A first objective of the present study was therefore to demonstrate that the design features of the BoDyS entail satisfactory interrater agreement rates and alpha reliability coefficients. At present, reliability analyses are confined to between-raters agreements; intrarater reliability still needs to be assessed.

Method

The two studies reported in this article were performed according to Declaration of Helsinki guidelines. Ethical approval was obtained by the local ethics committees of the University of Munich, the University of Tübingen, and the Medical University of Graz/Austria.

Participants

Preliminary data on between-listeners agreements of BoDyS scores have already been published in a study involving 25 individuals with dysarthria, predominantly with cerebrovascular accidents and head injuries (Nicola et al., 2004), and in a further study involving 22 individuals with cerebral palsy and dysarthria (Schölderle et al., 2016). In the present study, we collapsed the data from these two studies with interrater-agreement data obtained from two further studies: one involving 15 individuals with various etiologies, including Parkinson's disease, multiple sclerosis, encephalitis, head trauma, and stroke (Marten, 2008; Richter, 2008), and a further reliability study involving eight out of 48 individuals with Huntington's disease (Schölderle, 2008; Vögele, 2008). The reliability sample thus comprised 70 adults (28 women, 42 men) with dysarthria (22 cerebral palsy, 19 stroke, eight Huntington's disease, six head trauma, 15 others). All participants were older than 18 years, spoke German as their first language, and had a diagnosis of a neurological disorder. All participants had dysarthria according to the referring clinician; none of them were aphasic. Audio or video recordings of high quality were available in all cases. The overall severity of dysarthric impairment of the participants, as measured by the BoDyS total raw score, ranged between 0.6 and 3.9 (median: 2.8), which corresponds well with the overall distribution of severity values in the validation sample described later (see Table 5).

Assessment Procedure and Listeners

All assessments were performed according to BoDyS standards (Ziegler et al., 2015) and were audio- or video-recorded, depending on whether the individual signed an extra agreement form for video-recording. All recordings were evaluated by trained BoDyS raters according to the provisions sketched earlier. All raters had an education as speech-language therapists. Listener training involved familiarizing the trainees with the BoDyS scales and features and demonstrating multiple video-recorded examples of the 29 features in varying degrees and manifestations, with a joint elaboration of a consensus score for the associated BoDyS scale. Thereafter, the trainees completed evaluations of several video-recorded BoDyS examinations under the supervision of an experienced BoDyS user before they could be enrolled in the study.

Pooling across the four reliability studies mentioned earlier, a total of nine raters took part in the interrater-agreement analyses. They were grouped into six listener pairs who completed the analyses independently. Because several individuals from the reliability sample of 70 were assigned to more than one listener pair, 75 pairs of BoDyS examinations were entered into the reliability analyses overall. Collapsing across the nine scales and the 12 speech samples, a total of 7,407 pairs of itemwise BoDyS ratings —that is, 823 pairs per scale—were analyzed.[2]

Statistics

Reliability analyses were based on computations of Krippendorff's alpha using a routine developed by Hayes and Krippendorff (2007). The algorithm was implemented in SPSS (Version 23; IBM Corporation, 2015).

Results

Between-raters differences (absolute values) across the six listener pairs were calculated itemwise (i.e., single scores per scale and speech sample) and subjectwise (i.e., averaged scores across 12 items per subject; see Table 3). The entries in the left part of Table 3 reveal that at the single-item level, complete agreements between two raters (i.e., a difference of 0) occurred in 57% of all cases, with percentages ranging between 52% (MOD) and 61% (RSN) across the nine scales. As many as 95% of the overall between-raters differences were equal to or lower than 1. Regarding single scales, the percentages of agreement within 1 scale value ranged from 93% (TEM) to 98% (ART).

The right part of Table 3 lists agreement data for the mean scores across the 12 samples per subject. Recall that these data actually characterize the interrater reliability of the BoDyS profile. At the subject level, between 59% (VOQ, MOD) and 76% (RSP) of all differences were lower than 0.5 scale point on the BoDyS scale. More than 90% of all between-raters differences were within 1 scale point, and disagreement values greater than 1.2 occurred in only few cases, mostly on the TEM scale. Comparison of grand-average values across all scales and all items of each individual (BoDyS total score) revealed excellent interrater

Table 3. Interrater agreement: Percentages of between-raters differences less than or equal to thresholds in the top row.

	By item[a]				By subject[b]			
Scale	0	1	2	3	0.5	1.0	1.2	1.4
RSP	59	96	99	100	76	96	99	99
VOL	56	96	100	100	62	89	96	96
VOQ	54	97	100	100	59	89	97	97
VOS	57	94	99	100	70	91	94	97
ART	60	98	100	100	70	96	100	100
RSN	61	95	100	100	70	89	94	96
TEM	55	93	99	100	63	86	90	96
FLU	55	96	100	100	68	94	100	100
MOD	52	94	100	100	59	90	94	99
Pooled[c]	57	95	100	100	66	91	96	98
BoDyS total					96	100	100	100

Note. RSP = respiration; VOL = voice level; VOQ = voice quality; VOS = voice stability; ART = articulation; RSN = resonance; TEM = articulation rate; FLU = fluency; MOD = prosodic modulation.

[a]Single scores per speech sample and scale; 823 pairs per scale. [b]Average scores across the 12 BoDyS speech samples per scale—that is, actual scores of the BoDyS profile; 75 pairs per scale. [c]For "by item," 7,407 pairs across scales; for "by subject," 675 pairs across scales.

[2] Several of the individuals from the reliability sample had incomplete BoDyS protocols (i.e., were missing one or more test items or missing one of the three parallel versions). Therefore, the number of rating pairs was lower than 75 (individuals) × 12 (samples) × 9 (scales) = 8,100.

agreement (see bottom line of Table 3). On this measure of overall dysarthric impairment, 96% of all between-raters differences (72/75) were lower than 0.5 scale point, and no difference was greater than 0.73.

To determine a further statistical measure of scale reliability, Krippendorff's alpha (kalpha) was calculated itemwise and subjectwise for all listener pairs using an SPSS routine developed by Hayes and Krippendorff (2007). Note that reliability coefficients are required to estimate confidence intervals for standard scores as indices of measurement error. To control for agreement by chance, bootstraps on the basis of 5,000 samples (itemwise) and 2,500 samples (subjectwise), respectively, were used. Table 4 lists the results of these analyses. At the item level, kalpha values around .73 were obtained, with a minimum of .61 for the VOL scale and a maximum of .77 for the ART scale. Recall that these single-item agreement data do not reflect the ultimate reliabilities of the BoDyS scales. The subjectwise kalpha coefficients listed in the right part of Table 4, by contrast, constitute relevant reliability indices of the BoDyS profile. They assumed values between .76 (VOL) and .89 (RSP), with an average value of .85 across all scales and an average gain of .12 relative to the itemwise analyses.

All values in Table 4 had reasonably small 95% confidence intervals. Because a kalpha of less than .60 is considered to suggest questionable reliability, probabilities of failure to achieve a kalpha of at least .60 were calculated for all entries in Table 4 through bootstrapping (Hayes & Krippendorff, 2007). These probabilities were all lower than .001, except for the VOL scale in the itemwise case. Considering also that kalpha values around .80 are deemed to indicate good reliability, the reliabilities of the nine BoDyS variables averaged across the 12 speech samples per individual are more than satisfactory. As regards the BoDyS total score (see bottom line of Table 4), an excellent reliability index of .94 was obtained.

Discussion

In this study, interrater agreement was used as an index of test reliability. At the level of single items, exact agreement between two listeners occurred in 57% of all rating pairs across all scales, and agreement within 1 scale point in 95%. A comparison of these data with those reported by Bunton et al. (2007) must take into account the fact that the Mayo Clinic rating system uses 7-point scales as opposed to the 5-point BoDyS scales. The rate of exact interrater agreements was higher by almost 10% in the BoDyS than in the Mayo Clinic rating system. Gross between-raters deviations by an amount of at least one half of the scale length (i.e., 3 points or more in the Mayo Clinic system vs. 2 points or more in the BoDyS) occurred in 17% of the ratings reported by Bunton et al. (2007) but only 5% of the BoDyS ratings. This considerable advantage even at the single-item level is probably ascribable to the fact that the BoDyS scales dispense with the immense detail specified in the 38 Mayo Clinic scales. Another factor that may have been influential is the level of training of the raters. However, because one of the two groups of raters included in the Bunton et al. (2007) study consisted of highly experienced clinicians, this factor cannot explain the observed reliability differences.

Because the BoDyS profile is based on sum scores across the 12 test items for each scale rather than on individual item scores, a more realistic estimate of its reliability must be sought in between-listeners agreements on a

Table 4. Interrater agreement: Krippendorff's alpha (kalpha) for single scores (by item) and for sum scores across the 12 speech samples of each individual (by subject).

	By item[a]			By subject[b]		
Scale	kalpha	95% CI[c]	*p*(kalpha < .60)[d]	kalpha	95% CI	*p*(kalpha < .60)
RSP	.73	[.69, .78]	< .001	.89	[.84, .94]	< .001
VOL	.61	[.55, .67]	> .05	.76	[.67, .85]	< .001
VOQ	.73	[.69, .77]	< .001	.82	[.75, .88]	< .001
VOS	.69	[.64, .74]	< .001	.81	[.73, .89]	< .001
ART	.77	[.73, .80]	< .001	.88	[.84, .92]	< .001
RSN	.72	[.67, .76]	< .001	.87	[.80, .92]	< .001
TEM	.69	[.65, .74]	< .001	.79	[.71, .86]	< .001
FLU	.73	[.69, .77]	< .001	.86	[.81, .90]	< .001
MOD	.69	[.64, .73]	< .001	.81	[.73, .87]	< .001
Pooled[e]	.73	[.69, .76]	< .001	.85	[.83, .87]	< .001
BoDyS total				.94	[.91, .96]	< .001

Note. RSP = respiration; VOL = voice level; VOQ = voice quality; VOS = voice stability; ART = articulation; RSN = resonance; TEM = articulation rate; FLU = fluency; MOD = prosodic modulation.

[a] Single scores per speech sample and scale; 823 pairs per scale. [b] Data refer to the BoDyS scores ultimately used in clinical reports; 75 pairs per scale. [c] Computations of 95% confidence intervals are based on 5,000-sample bootstraps in the by-item analyses and 2,500-sample bootstraps in the by-subject analyses. [d] The .60 cutoff for kalpha was chosen because it is a threshold value indicating questionable reliability (Hayes & Krippendorff, 2007). [e] For "by item," 7,407 pairs across scales; for "by subject," 675 pairs across scales.

Table 5. Participant sample of validation study: Demographic data and BoDyS total scores.

Etiology	*N*	Gender (F/M)	Age (years)[a]	BoDyS total score[a,b]
Cerebrovascular accident	25	10/15	62 (48, 89)	3.2 (1.5, 3.7)
Closed head injury	12	5/7	23 (18,46)	2.4 (0.6, 3.2)
Cerebral palsy	27	9/18	23 (18, 56)	2.7 (1.4, 3.7)
Multiple sclerosis	16	9/7	51 (35, 73)	3.2 (2.2, 3.7)
Parkinson's disease	20	2/18	70 (36, 86)	3.2 (2.6, 3.9)
Progressive supranuclear palsy	21	7/14	69 (62, 81)	2.7 (0.9, 3.4)
Huntington's disease	25	11/14	46 (29, 68)	3.2 (2.0, 3.9)
Friedreich ataxia	16	8/8	39 (16, 71)	3.1 (2.4, 3.6)
Spinocerebellar ataxia (SCA3, 6)	16	11/5	54 (41, 80)	3.1 (2.6, 3.7)
Other	12	8/4	48 (20, 74)	3.0 (2.6. 3.7)
Total	190	80/110	52 (16, 89)	2.9 (0.6, 3.9)

Note. F = female; M = male.

[a] Median (minimum, maximum). [b] Grand average of raw scores across all items and all scales (range: 0–4).

by-subject rather than a by-item level. Across all scales, interrater agreements of these scores within 1 scale point were achieved at a rate of more than 90%, and deviations by 1.5 points were extremely rare. For the BoDyS total scores, calculated across all scales as a measure of overall severity, 96% of all rating pairs agreed within 0.5 scale interval, and no disagreements of more than 0.73 were observed. These data suggest that different raters, provided they are experienced in the administration of the BoDyS, will achieve reasonably good agreement in their evaluations of an individual's BoDyS profile.

Full standardization of a test, including estimates of measurement errors, requires computation of alpha reliability coefficients. Therefore, kalpha was determined for each of the nine scales and for the BoDyS total score. All coefficients were close to or higher than .80, which is interpretable as an indication of good reliability. The reliability coefficient of the BoDyS total score was excellent. Comparison of the by-subject with the by-item kalpha values revealed an average reliability gain of .12.

Among the overall highly acceptable reliability coefficients, those related to the TEM (i.e., articulation rate) and the VOL (i.e., voice level) scales ranged at the lower end. Hence, judgments of decreased or increased articulation rates (by neglecting the contribution of pauses or other disfluencies) and of vocal pitch and vocal loudness were obviously more sensitive to subjective variation than other scales. If circumstances call for a particularly high precision of measurements on these two variables, adjuvant acoustic measurements of articulation rate or fundamental frequency can be advised—for example, analyzing the recordings of the sentence-repetition tasks using an appropriate speech-wave analysis tool (for an application, see Brendel et al., 2015).

To summarize, the BoDyS is a reliable dysarthria test. Considering the provision of standard norms, an average reliability coefficient of around .85 across all scales yields sufficiently narrow confidence intervals—that is, measurement errors of a satisfactorily low magnitude. This result demonstrates that auditory speech evaluation may yield consistent measurements of dysarthria profiles. It still needs to be demonstrated, however, that equally high reliability coefficients can be achieved in an analysis of intrarater agreement. This will be one of the major goals in the future to complete our psychometric analyses of the BoDyS.

Construct Validity

Objectives

As outlined in the introduction, fuzzy test criteria and overlapping latent constructs may undermine not only the reliability but also the construct validity of the variables of a test. If dysarthria is assessed by a multivariate test, it has to be shown how the different variables combine to measure circumscribed underlying factors, and that different test variables measure different aspects of the speech impairment (discriminant validity). Moreover, if different assessment methods are used, it needs to be demonstrated that there is no interference between what is being assessed and how it is measured. This is important because changes in speech tasks may cause variations at different levels of the speech motor system in typical speakers (Tasko & McClean, 2004) and in individuals with dysarthria (Huber & Darling, 2011). The problem of mixed assessment methods is still more salient for assessments in which speech tasks are combined with nonspeech or paraspeech tasks or in which scales related to functional aspects (respiration, voice, etc.) are combined with scales related to communicative aspects (intelligibility, naturalness). Therefore, empirical data are required to examine if different assessment methods included in a test converge on the same construct (convergent validity).

The BoDyS differs from earlier clinical dysarthria assessment tools in that it conforms to a clearly structured test-theoretical model. It comprises nine latent factors measured by four elicitation methods, with three measurements per factor and method. The present study was performed to assess the discriminant and convergent validity of this design. Recall that discriminant validity might be challenged by the fact that the constructs supposedly measured by different BoDyS scales may interact on multiple levels, because the respiratory, laryngeal, and orofacial motor systems are intricately coupled at the neural-systems level (McClean & Tasko, 2002) and through biomechanical and aerodynamic interactions. Moreover, because prosodic impairments in dysarthria are a function of impairments at the respiratory, phonatory, and articulatory levels, performance on the TEM, FLU, and MOD scales necessarily depends on the integrity of the speech motor subsystems underlying the six remaining scales. Hence, a first objective of this part of the study was to examine the discriminant validity of the BoDyS scales—that is, to test if they actually measure different latent traits.

The second aim was to examine the convergent validity of the BoDyS scales. Because the BoDyS variables are assessed by different elicitation methods, one may ask whether these methods converge on the same latent factors. As an example, passage reading (READ) may be inherently more sensitive than sentence repetition (SENT) to disclosing problems with speech breathing or prosody, whereas the three voice scales may be less sensitive to this variation. This imbalance might distort the factor structure of the BoDyS, to the disadvantage of its construct validity.

The multitrait-multimethod (MTMM) technique proposed first by Campbell and Fiske (1959) provides a statistical tool to address questions concerning convergent and discriminant validity by a unified descriptive approach. Moreover, confirmatory factor analysis (CFA) permits direct testing of the BoDyS design by casting the BoDyS trait and method factors into a single measurement model. Application of this model to empirical data yields quantitative measures of the separate influences of traits and methods, and of the structural relationships between the nine traits and the four methods. This approach also allows for the testing of alternative hypotheses. Therefore, construct validation of the BoDyS presented here was based on a CFA modeling of its MTMM structure.

Methods

Participants

The validation study was based on a sample of 190 individuals with dysarthrias of different etiologies and varying degrees of severity. These individuals were drawn from a total of 339 German-speaking participants from 11 cohorts who had undergone BoDyS testing in different clinics (see the Appendix for a list of the cohorts included to date, and the collaborating clinics). Inclusion criteria for the validation sample were age above 16 years (postpuberty), German as first language, diagnosis of a neurological disorder, presence of dysarthria according to the referring clinician, no aphasic impairment, completed BoDyS testing (no missing speech samples), and availability of an audio or video recording of high quality. Table 5 lists demographic data of the validation sample by etiologic group, including a BoDyS total score calculated by averaging BoDyS raw scores across all items and all scales.

Assessment Procedure and Listeners

All assessments were performed according to BoDyS standards (Ziegler et al., 2015) and were audio- or video-recorded, depending on whether the individual signed an extra agreement form for video-recording. All recordings were evaluated by trained BoDyS raters, as described earlier. Across the whole evaluation of the BoDyS validation sample, 15 listeners were involved.

Statistics

MTMM Analysis

As mentioned earlier, the BoDyS scales are considered to represent nine distinct traits that are assessed using four different measurement methods: the four BoDyS elicitation modalities CONV, SENT, READ, and PICT. Major questions concerning the construct validity of the BoDyS are whether the scales actually assess different constructs (discriminant validity) and whether measurement of each construct is invariant with respect to the four measurement methods (convergent validity). Adequate answers to these questions are provided by the MTMM approach proposed by Campbell and Fiske (1959), which is based on an analysis of the correlations between the scores of all test variables in all measurement conditions. In our case, the MTMM correlation matrix contains the Pearson correlation coefficients between the scores of all BoDyS scales in all elicitation conditions—that is, a symmetric array of 9 × 4 rows and columns. All MTMM correlation analyses were implemented in SPSS (Version 23). In general, four classes of cells are distinguished in an MTMM matrix (Campbell & Fiske, 1959).

Monotrait-monomethod (MonT-MonM) cells. These cells constitute the main diagonal of the matrix. By convention, these cells are filled not by 1s (as is usually the case in symmetric matrices) but rather by the reliability coefficients of each scale in each modality, as measured by Cronbach's alpha. In the present case, reliability coefficients were determined across the three trials of each test modality. For each scale, four reliability coefficients were calculated—one for each modality. These coefficients represent the internal consistency of the BoDyS scales. Because a high consistency of the three scores of the same scale within the same modality is an essential requirement for test validity, the MonT-MonM coefficients are expected to be high. Because they furthermore reflect neither between-scales nor between-methods variation, they should attain the highest values in the MTMM matrix.

Monotrait-heteromethod (MonT-HetM) cells. These cells contain the Pearson correlation coefficients obtained from correlating the average score of each scale and modality with the average scores of the same scale in all other modalities, resulting in six MonT-HetM correlation coefficients per scale. The MonT-HetM coefficients of an MTMM matrix are considered an index of convergent validity. Under the assumption that the four BoDyS measurement methods converge upon the same trait for each scale, the MonT-HetM coefficients are expected to be high.

Heterotrait-monomethod (HetT-MonM) cells. These cells contain the correlation coefficients of pairs of scores from different scales within the same modality. In the BoDyS MTMM matrix, this results in 32 HetT-MonM coefficients. They are considered an index of discriminant validity. Under the assumption that the nine BoDyS scales measure different traits, their intercorrelations should be weak, even within the same elicitation method.

Heterotrait-heteromethod (HetT-HetM) cells. These cells contain all other correlation coefficients—that is, correlations between the average score of a scale in a certain modality with those of all other scales in all other modalities. There are 96 coefficients of this type for each scale in the MTMM matrix. Like the HetT-MonM coefficients, they measure the commonalities between different test variables, though across rather than within measurement methods, and are therefore considered a further index of discriminant validity. Because the HetT-HetM coefficients are sensitive to both between-traits and between-methods variation, they are expected to assume the lowest values in the MTMM matrix. The degree to which they are smaller than the HetT-MonM coefficients is considered an index of the influence of the methods factor.

CFA

One of the limitations of the MTMM approach we have described is that it involves only descriptive statistics, meaning that interpretations of the patterning of the MTMM matrix are debatable. Moreover, because the coefficients of the MTMM correlation matrix are spoiled by measurement errors, it is difficult to draw inferences about separate trait and method factors. CFA provides a means to model the effects of the underlying dimensions, separating the influences of the traits from those of the methods and from error variances. It thus allows for stringent statistical testing of the test-theoretical model underlying the BoDyS design, and permits inferential statistical testing of alternative models (Brown, 2015, Chapter 6). CFA modeling was carried out using the R package lavaan (Rosseel, 2012; http://lavaan.ugent.be/).

Results

The MTMM Matrix

Table 6 lists the mean, minimum, and maximum values of all four types of MTMM coefficients for each BoDyS scale. Inspection of the pooled results in the bottom line of Table 6 confirms that the BoDyS meets the criteria for convergent and discriminant validity almost perfectly. First, the reliability coefficients in the leftmost column (MonT-MonM) assumed the highest values in the matrix, with average α values between .90 and .95 across the nine scales and a grand average of .93, with reasonably small ranges of reliability coefficients. This result indicates that the BoDyS scores are highly consistent within scales and modalities. Second, the MonT-HetM coefficients were reasonably high as well, with an average of .82 and scalewise values between .69 (FLU) and .89 (RSN), suggesting that the scores within the same scale converge upon the same trait across the four different measurement methods (convergent validity). Third, the entries in the two rightmost columns (HetT-MonM and HetT-HetM) were considerably smaller than the MonT-HetM coefficients, with grand averages of .46 and .44, respectively. Of note, there was no overlap between the ranges of the coefficients indicating convergent validity (third column of Table 6) and indicating discriminant validity (fourth and fifth columns of Table 6), suggesting that the scales indeed represent different constructs. An exception was ART, which demonstrated relatively strong correlations with the MOD scale ($r = .70$) both within the SENT condition and between the SENT and the READ conditions, resulting in some overlap with the MonT-HetM correlations. Two further between-scales correlations were remarkably high—those between the TEM and MOD scales, both within the READ modality (.72) and between the READ and PICT modalities (.71). However, because the within-scale correlations of TEM and MOD were still substantially higher, the assumption of discriminant validity can still be maintained for these two scales. Fourth, the fact that the HetT-HetM coefficients were only slightly smaller than the HetT-MonM coefficients suggests that the (already weak) intercorrelations of the BoDyS scales were only in minor part ascribable to method effects.

CFA of the MTMM Matrix

To analyze the BoDyS MTMM matrix, a CFA measurement model was specified including four method factors (the four BoDyS elicitation methods) and nine trait factors

Table 6. Coefficients of the multitrait-multimethod matrix: Mean (minimum, maximum).

Scale	Monotrait-monomethod[a]	Monotrait-heteromethod[b]	Heterotrait-monomethod[c]	Heterotrait-heteromethod[d]
RSP	.93 (.89, .94)	.79 (.75, .85)	.50 (.33, .67)	.47 (.21, .65)
VOL	.93 (.92, .94)	.88 (.84, .92)	.37 (.18, .52)	.36 (.14, .59)
VOQ	.92 (.92, .92)	.87 (.83, .91)	.45 (.33, .53)	.44 (.29, .60)
VOS	.91 (.90, .93)	.81 (.72, .90)	.40 (.24, .61)	.39 (.17, .60)
ART	.95 (.93, .99)	.77 (**.61**, .87)	.49 (.25, **.70**)	.48 (.22, **.70**)
RSN	.95 (.94, .96)	.89 (.84, .94)	.42 (.24, .63)	.41 (.17, .68)
TEM	.94 (.94, .95)	.83 (.76, .87)	.50 (.25, **.72**)	.48 (.23, **.71**)
FLU	.90 (.88, .92)	.69 (.66, .77)	.44 (.18, .64)	.40 (.14, .61)
MOD	.93 (.92, .94)	.85 (.80, .89)	.53 (.37, **.72**)	.51 (.34, **.71**)
Total	.93 (.88, .99)	.82 (.76, .86)	.46 (.35, .58)	.44 (.33, .55)

Note. Boldface indicates a particularly high heterotrait or a particularly low monotrait correlation coefficient. RSP = respiration; VOL = voice level; VOQ = voice quality; VOS = voice stability; ART = articulation; RSN = resonance; TEM = articulation rate; FLU = fluency; MOD = prosodic modulation.

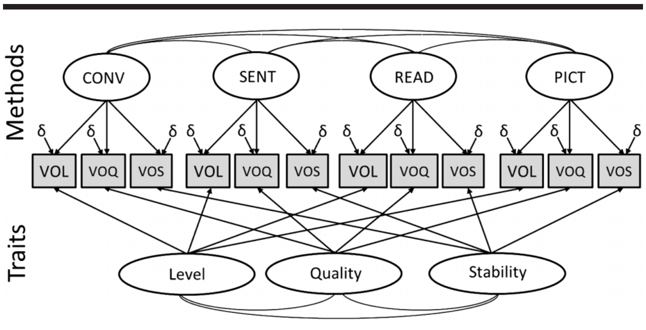
[a] Internal consistency (Cronbach's alpha; $n = 4$ per scale). [b] Convergent validity ($n = 6$ per scale). [c] Discriminant validity ($n = 32$ per scale). [d] Discriminant validity ($n = 96$ per scale).

(the nine BoDyS scales). The basic model assumption is that each BoDyS score is explainable as a linear function of the latent trait deemed to underlie the respective scale, the measurement (i.e., elicitation) method, and a measurement error δ. Figure 1 displays the path diagram of a section of this model, confined to only the three BoDyS voice scales VOL, VOQ, and VOS. In the full model, the trait-factor layer at the bottom comprises nine instead of only three factors, and the layer of indicator variables in the middle layer comprises 36 instead of 12 variables (one for each scale and each elicitation method).

The loadings of the indicator variables on their associated trait and method factors are represented by straight arrows. Each indicator is modeled to load on only one method and one trait factor. For instance, the VOL_{PICT} variable is modeled to load on the VOL trait factor and the PICT method factor. Once estimated, the factor loadings quantify the influences of the respective factors on their associated indicators. Convergent validity would entail that each indicator reflects predominantly the influence of the underlying latent trait, and much less that of the measurement method. Hence, the estimated factor loadings of the trait factors are expected to be considerably greater than those of the method factors.

The specifications of the CFA model applied here permitted the four method factors and the nine trait factors, respectively, to be correlated (curved lines in Figure 1), whereas cross-correlations between traits and methods were set to 0. Under the assumption of construct validity the method factors should be strongly intercorrelated (convergent validity), whereas correlations between trait factors (scales) should be relatively low (discriminant validity).

Figure 1. Confirmatory factor analysis model of a section of the BoDyS multitrait-multimodel matrix (path diagram). Only the three voice scales are displayed. Rectangles represent indicator variables—that is, average scores of each scale across the three trials within the respective modality. Note that for simplification purposes, the indicator variable names are not adjusted to their respective modalities. For instance, the term VOL represents different indicators depending on the dominating method factor—for example, VOL_{CONV}, VOL_{SENT}. Ellipses represent the latent factors considered to influence these scores—that is, the latent traits deemed to underlie the three voice scales (traits) and the four factors representing the elicitation methods (methods). The arrows designated by δ represent measurement errors. CONV = conversational speech; SENT = sentence repetition; READ = passage reading; PICT = picture-story narration; VOL = voice level; VOQ = voice quality; VOS = voice stability. For conventions, see Brown (2015, figure 6.1).



The specified model contained 13 latent factors (nine trait factors, four method factors) and 36 indicator variables. On the basis of the available data, the knowns in the input matrix of the model consisted of 36 indicator variances and 630 between-indicators covariances. The freely estimated model parameters consisted of 36 loadings each on the trait and method factors, 36 error variances, six covariances between the four method factors, and 36 covariances between the nine trait factors. Hence, the model was overidentified with 516 degrees of freedom. Estimation of the input MTMM matrix by this model used a maximum-likelihood minimization algorithm (Rosseel, 2012). The computations converged after 179 iterations. All freely estimated parameters were statistically significant. The model fit was good according to conventional fit criteria—that is, a low standardized root-mean-square residual (.049), a low root-mean-square error of approximation (.065), and a high comparative fit index (.955).[3]

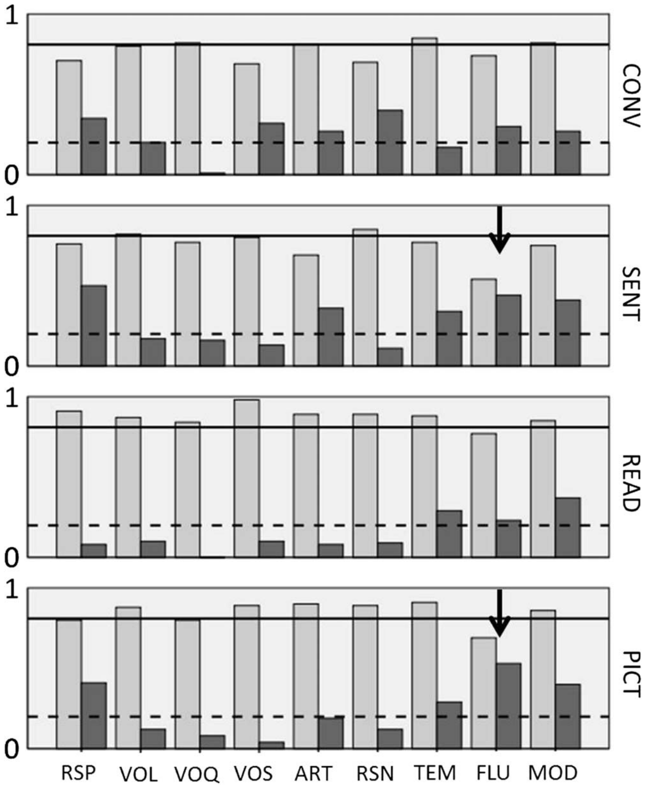
Figure 2 plots completely standardized loadings of the 36 indicator variables on the trait factors (light gray) and the method factors (dark gray). Grand averages of the trait- and method-factor loadings were λ_T = .81 and λ_M = .23, respectively (horizontal lines in Figure 2), confirming that the indicators were predominantly influenced by the latent traits underlying the nine scales, whereas the influences of the measurement methods were much smaller. Because the plotted factor loadings are completely standardized, squaring a factor loading provides the proportion of variance in the indicator explained by that factor. Hence, the average variance explained by the latent traits was $.81^2$—that is, 66%—whereas the variance explained by the methods was only $.23^2$—that is, 5%.

A more detailed inspection of the factor loadings depicted in Figure 2 reveals a so-called localized area of strain for the FLU_{SENT} indicator: Although this variable loaded relatively low on the trait factor associated with the FLU scale (λ = .54, corresponding to 29% of the variance), it loaded comparatively high on the SENT factor (λ = .44, corresponding to 19% of the variance). A similar situation, though less pronounced, was observed for the FLU_{PICT} indicator (.69 vs. .53, corresponding with 48% vs. 28%), suggesting that the BoDyS fluency scores are particularly sensitive to differences between elicitation methods.

Tables 7 and 8 report correlations between the latent-factor estimates of the CFA model. The correlations between the method factors were high, with an average of .76 (see Table 7), which is consistent with the assumption of convergent validity. The READ factor was an exception, because

[3] According to generally acknowledged criteria, the assumption of good model fit is fulfilled if the standardized root-mean-square residual is close to or smaller than .08, the root-mean-square error of approximation is close to or smaller than .06, the comparative fit index is greater than .95, and chi-square is lower than twice the number of degrees of freedom (Hu & Bentler, 1999).

Figure 2. Trait- (light gray) and method-factor loadings (dark gray) on the 36 indicator variables of the BoDyS in the confirmatory factor analysis model outlined in Figure 1. The loadings on variables pertaining to different elicitation methods are arranged in different panels (CONV: conversational speech; SENT: sentence repetition; READ: passage reading; PICT: picture-story narration). Grand averages of the trait- and method-factor loadings are indicated by solid and dashed lines, respectively. Recall that each indicator variable represents average scores across three parallel tasks. Areas of localized strain are indicated by arrows. RSP = respiration; VOL = voice level; VOQ = voice quality; VOS = voice stability; ART = articulation; RSN = resonance; TEM = articulation rate; FLU = fluency; MOD = prosodic modulation.



it showed a somewhat weaker relationship with the other method factors, especially with CONV, suggesting that speech elicitation through interview questions and passage reading may have differential effects on the BoDyS scores.

In comparison with the method factors, correlations between the trait factors were much lower, with an average of only .50 (see Table 8), indicating that the nine scales indeed measure different aspects of the dysarthria (discriminant validity). Again, there were outliers, especially concerning the correlations of the ART-, TEM-, and MOD-factor estimates of the model, with coefficients around .70 (in bold in Table 8). This suggests that judgments on the ART, TEM, and MOD scales share a comparatively large amount of common variance.

Table 7. Pearson correlation matrix for the method factors of the BoDyS confirmatory factor analysis model.

Method factor	CONV	SENT	READ
SENT	.86	—	
READ	.54	.71	—
PICT	.93	.85	.69

Note. CONV = conversational speech; SENT = sentence repetition; READ = passage reading; PICT = picture-story narration.

Table 8. Pearson correlation matrix for the trait factors of the BoDyS confirmatory factor analysis model.

Trait factor	RSP	VOL	VOQ	VOS	ART	RSN	TEM	FLU
VOL	.48	—						
VOQ	.55	.51	—					
VOS	.46	.32	.53	—				
ART	.65	.49	.54	.50	—			
RSN	.53	.38	.42	.28	.60	—		
TEM	.51	.35	.51	.60	.64	.42	—	
FLU	.64	.21	.37	.47	.54	.37	**.69**	—
MOD	.63	.49	.47	.49	**.73**	.46	**.74**	.61

Note. Boldface indicates outlier high correlation coefficients (see text). RSP = respiration; VOL = voice level; VOQ = voice quality; VOS = voice stability; ART = articulation; RSN = resonance; TEM = articulation rate; FLU = fluency; MOD = prosodic modulation.

Testing the Full CFA Model Against Alternative Models

One of the advantages of CFA modeling is that it permits hypothesis testing by comparing alternative models of the same data set against each other. This potential was exploited here to resolve some of the validity issues raised in the foregoing. For this purpose, the full model M_0 described in the preceding section (the parent model) was respecified in different ways by fixing the covariances of selected factors to 0 or 1 (nested models). This allowed us to simulate complete independence (covariance 0) or complete interdependence (covariance 1) between all or several trait factors and between the method factors. The fit indices of the resulting models are listed in Table 9. In the first line of this table, the fit indices of the parent model M_0 described earlier are listed for comparison. As a first step, the interdependence of the trait factors in the parent model M_0 was examined more closely.

Model M_I: Independent traits. This model postulates the null hypothesis that the latent traits measured by the nine BoDyS scales share no common variance—that is, all trait factors are uncorrelated (orthogonal traits). To simulate this assumption, the parent model M_0 was respecified by setting the correlations between all trait factors to 0. The resulting model M_1 was identified with 552 degrees of freedom and converged to a solution whose fit indices are listed in line 2 of Table 9. The figures reveal that the fit of M_1 was only marginally acceptable. A chi-square difference test of M_1 against M_0 suggested rejection of the assumption that the BoDyS scales measure completely independent constructs.

Table 9. Fit indices of the full confirmatory factor analysis model M_0 and five alternative nested models.

Model	χ^2 (*df*)	SRMR	RMSEA	CFI	χ^2_{diff} test[a]
M_0: parent model	923 (516)	.049	.065	.955	
M_1: independent traits[b]	1281 (552)	.087	.084	.919	$p < .001$
M_2: unique trait[c]	4798 (552)	.110	.202	.530	$p < .001$
M_3: unique ART/TEM/MOD trait[b]	1556 (519)	.063	.103	.885	$p < .001$
M_4: independent methods[b]	1069 (522)	.061	.074	.939	$p < .001$
M_5: unique method[c]	1012 (522)	.038	.071	.946	$p < .001$

Note. Goodness-of-fit criteria according to Hu and Bentler (1999) and Brown (2015, pp. 67–76): $\chi^2 < 2 \times df$, SRMR < .08, RMSEA < .06, CFI > .95. SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation; CFI = comparative fit index; ART = articulation; TEM = articulation rate; MOD = prosodic modulation.

[a]Significance level of χ^2 difference test of the model against M_0. [b]Between-factors correlations set to 0. [c]Between-factors correlations set to 1.

Note that this result does not disprove the discriminant validity of the BoDyS scales, because orthogonality is not a necessary requirement of discriminant validity. To the contrary, the assumption that the dysarthria signs measured by the nine BoDyS scales are completely independent is rather unlikely, given the many interactions between motor subsystems at the neural and functional levels. Hence, rejection of M_1 versus M_0 confirms that the traits measured by the nine scales, although their interrelations are relatively weak, are nonetheless subordinate to a common overarching theoretical construct—that is, dysarthria.

Model M_2: Unique trait. This model postulates a factor structure opposite that of M_1—that is, that all BoDyS scales measure the same construct. It represents a maximally restrictive null hypothesis against discriminant validity of the BoDyS profile, characterizing dysarthria as a nonseparable trait. To simulate this assumption, the parent model M_0 was respecified by setting the correlations between all trait factors to 1. The resulting model M_2 was identified with 552 degrees of freedom and converged to a solution whose fit indices are listed in line 3 of Table 9. The figures reveal that model M_2 fitted the data very poorly. Moreover, M_2 was significantly worse than M_0 according to a chi-square difference test. Hence, the assumption that the nine BoDyS scales measure the same construct can be rejected.

Model M_3: Unique ART/TEM/MOD trait. This model was designed to test a weaker null hypothesis against discriminant validity by assuming that the theoretical constructs deemed to underlie the ART, TEM, and MOD scales are identical. This hypothesis emerged from the finding of localized areas of strain in the correlations reported in the MTMM matrix and in Table 8. In model M_3, all covariances between the ART-, TEM-, and MOD-factor estimates were constrained to 1. Model M_3 was identified with 519 degrees of freedom. The fit indices listed in line 4 of Table 9 suggest poor model fit, especially due to high values for root-mean-square error of approximation and low values for comparative fit index. A chi-square difference test of M_3 versus M_0 led to rejection of this model. As an interpretation of this result, the outliers highlighted in Table 8 do not entail rejection of the BoDyS model distinguishing between ART, TEM, and MOD as distinct latent traits.

Further alternative models of nondiscriminant traits. Analogous to model M_3, a series of other models was designed by constraining pairs or trios of trait factors by complete collinearity, for the purpose of testing other specific assumptions of nonindependence of the nine BoDyS traits. As an example, concerns might be raised regarding the existence of three distinct traits underlying the prosody scales TEM, FLU, and MOD. None of the tested alternative models converged,[4] unfortunately, implying that further specific questions of local discriminant validity could not be answered by the statistical methods used here.

As a final step, the relevance of the method effects in M_0 was tested by specifying two further alternative models.

Model M_4: Independent methods. Model M_4 postulates that the four methods measure different constructs on each trait, as a strong null hypothesis against convergent validity of the BoDyS. Comparison of this "uncorrelated methods" model with the parent model M_0 yields a statistical test of whether the effects associated with the four BoDyS elicitation methods converge on the same traits. M_4 was specified by fixing all method-factor covariances to 0. The model was identified with 522 degrees of freedom and converged to a solution whose fit indices are listed in line 5 of Table 9. The results suggest that this model fitted the data with only marginally acceptable goodness. A chi-square difference test of M_4 versus M_0 led to rejection of this model.

Model M_5: Unique method. This last model postulates that the four BoDyS elicitation methods do not have any differential effect at all on the measurement of the latent traits. This model was simulated by fixing all between-methods covariances in M_0 to 1. The model was identified with 522 degrees of freedom and converged to a solution whose fit indices are listed in the bottom line of Table 9. The results suggest that this model fitted the data almost equally well as the full model M_0, in fact much better than any of the other alternative models discussed so far. Nonetheless,

[4] Nonconvergence of CFA models of MTMM matrices with lower dimensionalities is a common problem. It can be circumvented by using CFA models with correlated errors, but such models do not permit separate estimation of trait- and method-factor variances (Brown, 2015, Chapter 6).

there was still a significant chi-square difference between M_0 and M_5, demonstrating a superiority of the parent model and justifying the separate modeling of method factors in M_0. This result suggests that although the four elicitation methods do not do exactly the same thing, they converge very strongly in measuring the nine BoDyS traits (convergent validity).

Discussion

The BoDyS has obvious face validity, because the scales directly examine the speaking abilities of a person with dysarthria and because they are based on information that is, at least indirectly, amenable to the ears of the individual's interlocutors. Earlier studies have also confirmed several aspects of criterion validity of the BoDyS: Linear combinations of BoDyS scales permit rather accurate and plausible predictions of intelligibility measures and of naturalness ratings (Brendel et al., 2013; Schölderle et al., 2016), self-reports from people with dysarthria (Hinterberger et al., 2008), and laypersons' attitudes toward speakers with dysarthria (Schölderle et al., 2017). Hence, the BoDyS profiles represent meaningful information about an individual's speech impairment. Overall, the test fulfills important requirements of face and criterion validity.

The present study was focused on another aspect of validity: the questions of whether the nine BoDyS scales measure different theoretical constructs (discriminant validity) and whether their assessment is invariant under the four different elicitation methods of the BoDyS protocol (convergent validity). An ensuing issue was how the conceived latent factors underlying the BoDyS scales and the four elicitation methods are related to each other.

Convergent and Discriminant Validity as Evidenced by the MTMM Matrix

The MTMM approach proposed by Campbell and Fiske (1959) was chosen to inspect the array of correlations within and between traits and methods. The findings were strongly in favor of convergent validity, because high internal consistency and high MonT-HetM correlations were observed, demonstrating that the multiple trials within each elicitation modality and the different elicitation methods converged on the same constructs. Furthermore, inspection of the MTMM matrix also supported the assumption of discriminant validity, because the correlations between traits—both within and across methods—were much weaker than the correlations between different methods within the same trait.

Convergent and Discriminant Validity as Evidenced by CFA Modeling

After we cast the trait and method factors of the BoDyS test structure in a CFA measurement model with correlated traits and correlated methods, the assumptions of construct validity were corroborated still more convincingly. First, the statistical model of the trait- and method-factor design of the BoDyS fitted the empirical data with considerable accuracy. Second, the indicator variables of the model—that is, the average scores per trait and method —loaded high on the trait factors and considerably lower on the method factors, confirming that the test scores were predominantly influenced by the traits and much less so by the methods. The CFA model thus permitted separation of the variance proportions explained by the latent factors underlying the nine scales and those underlying the four methods. Whereas the trait variance amounted to 66% on average, the method factors contributed only 5%, yielding a convincing quantitative account of the convergent validity of the BoDyS. That is, convergent validity of the BoDyS was supported by the fact that information from the nine scales contributed much more strongly to the data than did the methods used to elicit the information. Third, the correlations between the latent method-factor estimates were considerably high, confirming convergence of the four BoDyS elicitation methods. In contrast, the latent trait-factor estimates had much lower correlations, indicating discriminant validity.

Testing Hypotheses Against Convergent and Discriminant Validity

Testing of this CFA model against two standard alternative models, M_1 and M_2, revealed that neither a hypothesis of complete independence of the nine trait factors nor one of a unique underlying trait fitted the empirical data of this study to an acceptable extent, and both alternatives could be rejected. Hence, there is common variance between the latent factors represented by the nine BoDyS scales, suggesting that they pertain to a common overarching trait—that is, dysarthria. Nonetheless, the subdivision of this underlying construct into separate scales in the BoDyS factor design was corroborated convincingly. Further attempts at testing the full CFA model against hypotheses postulating uniqueness of pairs or trios of trait-factor estimates, regrettably, were only partly successful, because most of these alternative models failed to converge. Only model M_3, relating to one of the more conspicuous sections of the MTMM matrix—that is, the relationship between the ART, TEM, and MOD factors—converged. Analysis of M_3 showed that combining these factors into a unique underlying trait did not yield an acceptable fit to the empirical data. Despite their comparably strong interrelations in the correlation matrix of the factor estimates (see Table 8), the distinction of three separate ART, TEM, and MOD factors in the BoDyS design is still warranted.

Last, two further models (M_4 and M_5) were designed to test alternative hypotheses concerning the influence of the BoDyS method factors. The assumption of a unique method factor, simulated by fixing the model-factor covariances to 1 in M_5, fitted the empirical data more closely than any of the other alternative models tested, demonstrating that the four elicitation methods influence the BoDyS profile in essentially the same manner. This result is a strong statistical proof of the convergent validity of the BoDyS. Yet M_5 was still weaker than M_0, demonstrating that the design feature of using different elicitation methods in fact adds a source of significant variation to the BoDyS measurement model without destroying its latent-trait structure.

Conspicuous Patterns in the BoDyS Factor Structure

Though the available statistics provide comprehensive and very conclusive evidence in favor of the construct validity of the BoDyS, some of the findings presented in Tables 7 and 8 were conspicuous of local strains or weaknesses of the construed test factor structure. Yet although these minor structural flaws may cause traces of concern from a formal perspective, they are highly trackable from a theoretical and clinical viewpoint. We find it remarkable that they predominantly involve interactions with the prosodic factors TEM, FLU, and MOD. Correlations higher than .60 occurred, for instance, between these factors and the RSP and ART factors, reflecting the obvious dependence of prosodic patterns on the integrity of the motor subsystems of speaking. Though these interdependencies may derogate a pure concept of discriminant validity, they are theoretically highly transparent and would not, from a clinical perspective, justify collapsing any of the respiratory, voice, or articulation and resonance scales with one of the prosodic scales, or omitting one of them.

Regarding the correlation matrix of the method-factor estimates (see Table 7), one coefficient was conspicuously low: CONV versus READ. This is a remarkable result, considering that some approaches (e.g., the Mayo Clinic rating system) strongly favor passage reading to elicit speech samples. The elicitation methods representing the CONV and READ factors vary along two important dimensions: On the one hand, passage reading leaves fewer options than conversational speech to control utterance length and may therefore reveal respiratory or prosodic problems and fatigue, if present. On the other hand, answering interview questions may, to a greater extent than passage reading, deplete cognitive planning resources, at the expense of the resources a person with dysarthria needs to maintain proper articulation, voice level, or articulation rate. Due to the partly complementary properties of these two elicitation methods, they may disclose differential aspects of the speech impairments of individuals with dysarthria without destroying the factor composition underlying the BoDyS scales.

Conclusions and Implications

The Bogenhausen Dysarthria Scales is a standardized test for the clinical evaluation of dysarthria. The assessment is based on auditory-perceptual evaluations of speech samples by clinical experts. Unlike the FDA-2 (Enderby & Palmer, 2008) or other composite dysarthria examination instruments, the BoDyS is homogeneous as regards its exclusive focus on speech tasks as its test items and on auditory-perceptual ratings as its measurement method. Compared with the Mayo Clinic rating system, the BoDyS is based on a much smaller number of scales (nine vs. 38) and a much larger number of items per scale (12 vs. one). This design was chosen to optimize the reliability and validity of the test. A major property distinguishing the BoDyS from other dysarthria assessment tools is that its design rests on a test-theoretical model that allows for an in-depth analysis of its construct validity.

The BoDyS have already been used in a number of investigations of dysarthria in different etiologic groups and have undergone several preliminary studies of interrater agreement and criterion validity (Brendel et al., 2013, 2015; Schölderle et al., 2013, 2016, 2017). A collection of this and other pertinent work is listed in the Appendix.

Due to its clear and theoretically transparent arrangement of assessment scales and the statistical safeguarding of test scores through multiple ratings, the BoDyS is reliable in terms of interrater agreement. The magnitudes of the alpha reliability coefficients of the BoDyS scales reported in this article guarantee narrow confidence intervals of standardized test scores, assuming appropriate rater training.

The BoDyS also possess high discriminant and convergent validity. The four elicitation methods converge on the same latent test variables, but nonetheless introduce some interpretable and diagnostically useful variability. The trait factors are sufficiently independent to justify a design that is based on nine scales. At the same time, there is clinically and theoretically tractable covariation between the scales, predominantly due to the obvious dependency of the three prosodic factors on the factors representing respiratory, phonatory, and articulatory functioning. Yet between-traits covariation probably also results from third-variable influences (severity of neurologic impairment, lesion size), from the neural integration of speech motor subsystems (McClean & Tasko, 2002), and from mechanical and aerodynamic interactions between subsystems.

The BoDyS is outstanding among the dysarthria assessments published so far because standard norms are available for all scales. As a consequence, the test is useful to measure between-subjects differences of test profiles, longitudinal changes of speech impairment, and differences between the nine test variables. The test may therefore serve as a core standard instrument in the clinical assessment of dysarthria and as a measurement tool in dysarthria research. As an expansion, the present/absent ratings of the 29 BoDyS features can be used for the purpose of obtaining therapeutically relevant information or classifying dysarthria subtypes (Schölderle et al., 2013), although such inferences are not supported by psychometric evidence. A further expansion by specific acoustic parameters is advisable if greater accuracy is required on certain dimensions—such as articulation rate, pause durations, or voice fundamental frequency—in research. The BoDyS speech samples lend themselves to such analyses using standard speech-wave analysis software (for an application, see Brendel et al., 2013).

It should be reiterated that use of the BoDyS—as with any other standard test—does not preclude clinicians from performing further individualized examinations for the purpose of obtaining more specific diagnostic information or planning therapeutic interventions. Moreover, administration of the BoDyS does not replace assessments of communication-related parameters, especially intelligibility. However, the BoDyS provides a psychometrically sound diagnostic basis upon which more exploratory assessments or examinations of levels of communicative activity and participation of a person with dysarthria can build. Not least, application of standardized diagnostic tools is a key prerequisite for the planning and implementation of research programs and treatment trials of dysarthria in the future.

Acknowledgments

This research was supported by German Research Association Grants ZI469/10-1/10-2/10-3 and ZI469/15-1/15-2 (awarded to Wolfram Ziegler), and by a scholarship of the German Academic Scholarship Foundation (to Theresa Schölderle). We are grateful to ReHa-Hilfe for continuous support, and to the many institutions, clinicians, and students who contributed to this work. We also wish to thank the individuals with dysarthria who participated in the BoDyS projects.

References

Auzou, P., & Rolland-Monnoury, V. (2006). *Batterie d'Évaluation Clinique de la Dysarthrie* [*Clinical Dysarthria Evaluation Battery*]. Isbergues, France: Ortho Édition.

Brendel, B., Ackermann, H., Berg, D., Lindig, T., Schölderle, T., Schöls, L., … Ziegler, W. (2013). Friedreich ataxia: Dysarthria profile and clinical data. *The Cerebellum, 12,* 475–484.

Brendel, B., Synofzik, M., Ackermann, H., Lindig, T., Schölderle, T., Schöls, L., & Ziegler, W. (2015). Comparing speech characteristics in spinocerebellar ataxias type 3 and type 6 with Friedreich ataxia. *Journal of Neurology, 262,* 21–26.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford.

Bunton, K., Kent, R. D., Duffy, J. R., Rosenbek, J. C., & Kent, J. F. (2007). Listener agreement for auditory-perceptual ratings of dysarthria. *Journal of Speech, Language, and Hearing Research, 50,* 1481–1495.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Chenery, H. J. (1998). Perceptual analysis of dysarthric speech. In B. E. Murdoch (Ed.), *Dysarthria. A physiological approach to assessment and treatment* (pp. 36–67). Cheltenham, United Kingdom: Stanley Thornes.

Cohen, R. J., Swerdlik, M. E., & Sturman, E. D. (2013). *Psychological testing and assessment: An introduction to tests and measurement* (8th ed.). New York, NY: McGraw-Hill.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1975). *Motor speech disorders*. Philadelphia, PA: Saunders.

Derdemezis, E., Vorperian, H. K., Kent, R. D., Fourakis, M., Reinicke, E. L., & Bolt, D. M. (2016). Optimizing vowel formant measurements in four acoustic analysis systems for diverse speaker groups. *American Journal of Speech-Language Pathology, 25,* 335–354.

Duffy, J. R. (2013). *Motor speech disorders: Substrates, differential diagnosis, and management* (3rd ed.). St. Louis, MO: Elsevier Mosby.

Duffy, J. R., & Kent, R. D. (2001). Darley's contributions to the understanding, differential diagnosis, and scientific study of the dysarthrias. *Aphasiology, 15,* 275–289.

Enderby, P., & Palmer, R. (2008). *FDA-2: Frenchay Dysarthria Assessment–Second Edition Examiner's Manual*. Austin, TX: Pro-ed.

Enderby, P., & Palmer, R. (2012). *FDA-2: Frenchay Dysarthrie Assessment–2*. Idstein, Germany: Schulz-Kirchner Verlag.

Forrest, K., & Weismer, G. (2009). Acoustic analysis of motor speech disorders. In M. R. McNeil (Ed.), *Clinical management of sensorimotor speech disorders* (2nd ed., pp. 46–63). New York, NY: Thieme.

Ganz, S. (2004). *Dimensionen der Dysarthriediagnostik: Selbsteinschätzung durch die Betroffenen - Schätzurteile anhand der Bogenhausener Dysarthrieskalen - Spezifische Untersuchungen zur Sprechatmung* [*Dimensions of dysarthria assessment: Patient self reports - BoDyS ratings - pecific investigations of speech breathing*] (Master's thesis). Ludwig-Maximilians-Universität, München, Germany.

Glauner, F. (2015). *Wie beurteilen sprachtherapeutische Laienhörer die Sprechnatürlichkeit bei ataktischen Dysarthrien? Eine Untersuchung bei hereditären Ataxien* [*How do laypersons judge speech naturalness of ataxic dysarthria? A study of hereditary ataxia*] (BA-Thesis). Ludwig-Maximilians-Universität, München, Germany.

Hartelius, L., & Svensson, P. (1990). *Dysartritest*. Stockholm, Sweden: Psykologiförlaget.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1,* 77–89.

Heß, B. (2013). *"Diese Person ist sympathisch." Einstellungen von Laien gegenüber Menschen mit Dysarthrie bei infantiler Cerebralparese* [*"This person is likeable." Laypersons' attitudes towards persons with dysarthria in cerebral palsy*] (Master's thesis). Ludwig-Maximilians-Universität, München, Germany.

Hinterberger, K., Ostwald, A., Löper, M. L., Levin, J., Lorenzl, S., & Ziegler, W. (2008). Verlauf und Schweregrad der Dysarthrie bei Patienten mit progressiver supranukleärer Blickparese (PSP) und idiopathischem Parkinson-Syndrom (IPS) [*Course and severity of dysarthria in patients with progressive supranuclear palsy (PSP) and Parkinson's disease*]. *Neurologie & Rehabilitation, 14,* 247–253.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Huber, C. (2007). *Dysarthrie bei Morbus Huntington: Eine Untersuchung zur Verständlichkeit* [*Dysarthria in Huntington's disease: An investigation of intelligibility*] (BA-Thesis). Ludwig-Maximilians-Universität, München, Germany.

Huber, J. E., & Darling, M. (2011). Effect of Parkinson's disease on the production of structured and unstructured speaking tasks: Respiratory physiologic and linguistic considerations. *Journal of Speech, Language, and Hearing Research, 54,* 33–46.

IBM Corporation. (2015). IBM SPSS Statistics for Windows (Version 23) [Computer software]. Armonk, NY: IBM Corp.

Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology, 5*(3), 7–23.

Kent, R. D. (2009). Perceptual sensorimotor speech examination for motor speech disorders. In M. R. McNeil (Ed.), *Clinical management of sensorimotor speech disorders* (2nd ed., pp. 19–29). New York, NY: Thieme.

Kent, R. D. (2015). Nonspeech oral movements and oral motor disorders: A narrative review. *American Journal of Speech-Language Pathology, 24,* 763–789.

Kent, R. D., & Kim, Y.-J. (2003). Toward an acoustic typology of motor speech disorders. *Clinical Linguistics & Phonetics, 17,* 427–445.

Klabuschnig, M. (2007). *Chorea Huntington - Analyse von Dysarthrieprofilen mit Schwerpunkt auf prosodischen Störungen* [*Huntington's chorea - Analysis of dysarthria profiles with a focus on prosodic impairment*] (BA-Thesis). Ludwig-Maximilians-Universität, München, Germany.

Kleinow, J., & Smith, A. (2006). Potential interactions among linguistic, autonomic, and motor factors in speech. *Developmental Psychobiology, 48,* 275–287.

Körner, F. (2016). *Dysarthrie bei Multipler Sklerose: Störungsbilder und subjektive Beeinträchtigung* [*Dysarthria in Multiple Sclerosis: Clinical patterns and subjective impairment*] (Master's thesis). Ludwig-Maximilians-Universität, München, Germany.

Lansford, K. L., & Liss, J. M. (2014). Vowel acoustics in dysarthria: Speech disorder diagnosis and classification. *Journal of Speech, Language, and Hearing Research, 57,* 57–67.

Liss, J. M., Spitzer, S. M., Caviness, J. N., & Adler, C. (2002). The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria. *The Journal of the Acoustical Society of America, 112,* 3022–3030.

Löper, M. L. (2007). *Dysarthrie bei Parkinson-Syndromen: Ein Vergleich verschiedener Untersuchungsmethoden* [*Dysarthria in Parkinson's syndromes: A comparison of different assessment methods*] (BA-Thesis). Ludwig-Maximilians-Universität, München, Germany.

Mallien, G. (2011). *Explorative multizentrische Querschnittstudie zur Diagnostik der Dysarthrie bei Progressiver Supranukleärer Blickparese – PSP* [*Exploratory multicentre cross-sectional study of dysarthria assessment in progressive supranuclear palsy – PSP*] (PhD-Thesis). Universität Potsdam, Potsdam, Germany.

Marten, K. (2008). *Zum Zusammenhang von nonverbalen und verbalen Untersuchungsmethoden in der Dysarthriediagnostik: Ein Vergleich zwischen der Frenchay Dysarthrie-Untersuchung und den Bogenhausener Dysarthrieskalen (BoDyS)* [*Relationship between nonspeech and speech approaches in dysarthria assessment: A comparison between the Frenchay Dysarthria Assessment and the Bogenhausen Dysarthria Scales (BoDyS)*] (Diploma-Thesis). Universität Potsdam, Potsdam, Germany.

McClean, M. D., & Tasko, S. M. (2002). Association of orofacial with laryngeal and respiratory motor output during speech. *Experimental Brain Research, 146,* 481–489.

Nicola, F., Ziegler, W., & Vogel, M. (2004). Die Bogenhausener Dysarthrieskalen (*BoDyS*): Ein Instrument für die klinische Dysarthriediagnostik [*The Bogenhausen Dysarthria Scales (BoDyS): An instrument for clinical dysarthria assessment*]. *Forum Logopädie, 2*(18), 14–22.

Ostwald, A. (2007). *Dysarthrieprofile bei Patienten mit idiopathischem Parkinson-Syndrom* [*Dysarthria profiles in patients with idiopathic Parkinson's disease*] (BA-Thesis). Ludwig-Maximilians-Universität, München, Germany.

Pokorny, C. I. (2011). *Evaluation einer Light-Version der Bogenhausener Dysarthrieskalen* [*Evaluation of a light-version of the Bogenhausen Dysarthria Scales*] (Diploma-Thesis). RWTH Aachen University, Aachen, Germany.

Özsancak, C., Auzou, P., & Hannequin, D. (2001). Measurement of voice onset time in dysarthric patients: Methodological considerations. *Folia Phoniatrica et Logopaedica, 53,* 48–57.

Protti, D. (2014). *La valutazione della disartria, il BoDyS e il Robertson Profile; due approci a confronto* [*The evaluation of dysarthria, BoDyS, and the Robertson Profile; comparison of two approaches*] (BA-Thesis). Università degli Studi di Torino, Torino, Italy.

Richter, A. (2008). *Entwicklung von sprachlichem Diagnostikmaterial zur Prüfung der Resonanz und der Artikulation im Rahmen einer Dysarthrie* [*Development of verbal materials for the assessment of resonance and articulation in dysarthria*] (Diploma-Thesis). Universität Potsdam, Potsdam, Germany.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Sacher, A. (2016). *Wie sensitiv ist die Dysarthriediagnostik bei Morbus Parkinson? Eine Untersuchung zur Störungsausprägung in Abhängigkeit von Verlauf und Sprechmodalität* [*How sensitive is dysarthria assessment in Parkinson's disease? An investigation of severity as a function of the course of the disease and speaking modality*] (Master's thesis). Paris Lodron Universität, Salzburg, Austria.

Schmich, J. (2007). *Chorea Huntington. Anaylse von Dysarthrieprofilen mit Schwerpunkt auf Stimmstörungen* [*Huntington's chorea: Analysis of dysarthria profiles with a focus on voice impairments*] (BA-Thesis). Ludwig-Maximilians-Universität, München, Germany.

Schmich, J. (2009). *Kommunikations- und alltagsbezogene Diagnostik bei Dysarthrie-Patienten mit Basalganglienerkrankungen* [*Assessment related to communication and daily life of patients with Dysarthria in basal ganglia disorders*] (Master's thesis). Ludwig-Maximilians-Universität, München, Germany.

Schmid, S. (2012). *Sprechen im Alter. Vergleichende akustische Analyse von gesunden Sprechern und Patienten mit leichter Dysarthrie aus zwei Altersgruppen* [*Speaking in old age. Comparative acoustic analysis of typical speakers and patients with mild Dysarthria from two age groups*] (BA-Thesis). Ludwig-Maximilians-Universität, München, Germany.

Schmid, S. (2015). *Dysarthriediagnostik im Spannungsfeld zwischen Innen- und Außenperspektive* [*Dysarthria assessment between inner- and outside perspectives*] (Master's thesis). Ludwig-Maximilians-Universität München, Germany.

Schölderle, T. (2008). *Verständlichkeit und Natürlichkeit des Sprechens bei Patienten mit Chorea Huntington* [*Intelligibility and naturalness of speech in patients with Huntington's disease*] (BA-Thesis). Ludwig-Maximilians-Universität, Munich, Germany.

Schölderle, T. (2014). *The impact of early brain damage on speech: Features and characteristics of dysarthria in adults with cerebral palsy* (PhD-Thesis). Ludwig-Maximilians-Universität, München, Germany.

Schölderle, T., Staiger, A., Heß, B., & Ziegler, W. (2017). *The impact of dysarthria on laypersons' attitudes towards adults with cerebral palsy*. Manuscript submitted for publication.

Schölderle, T., Staiger, A., Lampe, R., Strecker, K., & Ziegler, W. (2016). Dysarthria in adults with cerebral palsy: Clinical presentation and impacts on communication. *Journal of Speech, Language, and Hearing Research, 59,* 216–229.

Schölderle, T., Staiger, A., Lampe, R., & Ziegler, W. (2013). Dysarthria syndromes in adult cerebral palsy. *Journal of Medical Speech-Language Pathology, 20,* 100–105.

Schwab, K. (2016). *Maximale Vokalhaltedauer als diagnostisches Maß für Sprechatmungsstörungen? Eine Untersuchung typischer Sprecher unterschiedlichen Alters* [*Maximum sustained vowel duration as a diagnostic measure of impaired speech breathing? An investigation of typical speakers of different ages*] (BA-Thesis). Ludwig-Maximilians-Universität, München, Germany.

Schwager, V. (2010). *Dysarthrie bei Patienten mit supranukleärer Blickparese (PSP). Verständlichkeit und Natürlichkeit beim*

Lesen und Nachsprechen [*Dysarthria in patients with progressive supranuclear palsy (PSP). Intelligibility and naturalness in text reading and repetition*] (Master's thesis). Ludwig-Maximilians-Universität, München, Germany.

Sedlaczek, A. (2016). *Einfluss von Alter und Geschlecht auf das Sprechtempo: Normdaten für ein Verfahren zur Dysarthriediagnostik* [*Influence of age and sex on speaking rate: Standard norms for a dysarthria assessment tool*] (BA-Thesis). Ludwig-Maximilians-Universität, München, Germany.

Sheard, C., Adams, R. D., & Davis, P. J. (1991). Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility. *Journal of Speech and Hearing Research, 34,* 285–293.

Simson, E. (2004). *Dimensionen der Dysarthriediagnostik: Selbsteinschätzung durch die Betroffenen - Schätzurteile anhand der Bogenhausener Dysarthrieskalen - Spezifische Untersuchungen zur Stimme* [*Dimensions of dysarthria assessment: Patient self reports - BoDyS ratings - Specific investigations of voice*] (Master's thesis). Ludwig-Maximilians-Universität, München, Germany.

Staiger, A., Schölderle, T., Brendel, B., Bötzel, K., & Ziegler, W. (2016). Oral motor abilities are task dependent: A factor analytic approach to performance rate. *Journal of Motor Behavior.* https://doi.org/10.1080/00222895.2016.1241747

Staiger, A., Schölderle, T., Brendel, B., & Ziegler, W. (2017). Dissociating oral motor capabilities: Evidence from patients with movement disorders. *Neuropsychologia, 95,* 40–53.

Steinbauer, K. M. (2011). *Dysarthrieprofile nach BoDyS: Ein Vergleich anhand dreier Krankheitsbilder* [*BoDyS dysarthria profiles: A comparison of three syndromes*] (Diploma-Thesis). RWTH Aachen University, Aachen, Germany.

Tasko, S. M., & McClean, M. D. (2004). Variations in articulatory movement with changes in speech task. *Journal of Speech, Language, and Hearing Research, 47,* 85–100.

Vögele, B. (2008). *Dysarthrieprofile bei Patienten mit Chorea Huntington* [*Dysarthria profiles in patients with Huntington's disease*] (BA-Thesis). Ludwig-Maximilians-Universität, Munich, Germany.

Wannberg, P., Schalling, E., & Hartelius, L. (2016). Perceptual assessment of dysarthria: Comparison of a general and a detailed assessment protocol. *Logopedics Phoniatrics Vocology, 41,* 159–167.

Weismer, G. (2006). Philosophy of research in motor speech disorders. *Clinical Linguistics & Phonetics, 20,* 315–349.

World Health Organization. (2001). *The International Classification of Functioning, Disability and Health*. Geneva, Switzerland: Author.

Ziegler, W. (2003). Speech motor control is task-specific: Evidence from dysarthria and apraxia of speech. *Aphasiology, 17,* 3–36.

Ziegler, W., & Ackermann, H. (2013). Neuromotor speech impairment: It's all in the talking. *Folia Phoniatrica et Logopaedica, 65,* 55–67.

Ziegler, W., & von Cramon, D. (1983). Vowel distortion in traumatic dysarthria: A formant study. *Phonetica, 40,* 63–78.

Ziegler, W., Schölderle, T., Staiger, A., & Vogel, M. (2015). Die Bogenhausener Dysarthrieskalen (BoDyS): Ein standardisierter Test für die Dysarthriediagnostik bei Erwachsenen [*The Bogenhausen Dysarthria Scales (BoDyS): A standardized test for the assessment of dysarthria in adults*]. *Sprache, Stimme, Gehör, 39,* 171–175.

Ziegler, W., & Zierdt, A. (2008). Telediagnostic assessment of intelligibility in dysarthria: A pilot-investigation of MVP-online. *Journal of Communication Disorders, 41,* 553–577.

Zimmermann, I. (2010). *Verständlichkeitsuntersuchungen bei Patienten mit Progressiver Supranukleärer Blickparese* [*Intelligibility assessment in patients with progressive supranuclear palsy*] (BA-Thesis). Universität Potsdam, Potsdam, Germany.

Zyski, B. J., & Weisiger, B. E. (1987). Identification of dysarthria types based on perceptual analysis. *Journal of Communication Disorders, 20,* 367–378.

Appendix

Overview of the Current State of Development of the BoDyS

Cohort	Etiologies	*N*	Collaborating departments	References
1	Stroke, head injury, others	25	Clinic for Neuropsychology, City Hospital Bogenhausen, Munich, Germany (G. Goldenberg)	Nicola, Ziegler, and Vogel (2004); Schmid (2015)
2	Stroke, head injury, others	20	Clinic for Neuropsychology, City Hospital Bogenhausen, Munich, Germany (G. Goldenberg)	Ganz (2004); Simson (2004)
3	Stroke, head injury, others	15	Center for Ambulant Rehabilitation, Berlin, Germany (D. Steube); Neurologic Clinic for Movement Disorders and Parkinsonism, Beelitz, Germany (G. Ebersbach)	Marten (2008); Richter (2008)
4	Huntington's disease	46	Medical University Clinic Graz, Austria (U. Mannsberger, R. Bonelli)	Huber (2007); Klabuschnig (2007); Pokorny (2011); Schmich (2007); Schölderle (2008); Steinbauer (2011); Vögele (2008)
5	Parkinson's syndromes	16	Clinic for Neurology, University of Munich, Germany (S. Lorenzl)	Hinterberger et al. (2008); Löper (2007); Ostwald (2007); Pokorny (2011); Schmich (2009); Steinbauer (2011)
6	Progressive supranuclear palsy	50	Clinic for Neurology, University of Munich, Germany (S. Lorenzl) Neurologic Clinic for Movement Disorders and Parkinsonism, Beelitz, Germany (G. Ebersbach)	Mallien (2011); Schwager (2010); Zimmermann (2010)
7	Hereditary ataxias	41	Department of General Neurology, University of Tübingen, Germany (L. Schöls, H. Ackermann)	Brendel et al. (2013, 2015); Glauner (2015); Pokorny (2011); Steinbauer (2011)
8	Cerebral palsy	22	Integration Center for Cerebral Palsy, Munich, Germany (R. Lampe)	Heß (2013); Schölderle (2014); Schölderle, Staiger, Lampe, Strecker, and Ziegler (2016); Schölderle, Staiger, Lampe, and Ziegler (2013)
9	Multiple sclerosis	21	Schmieder Clinics, Konstanz, Germany (C. Dettmers, B. Gröne)	Körner (2016)
10	Parkinson's disease	13	Clinic for Neurology, University of Regensburg, Germany (U. Bogdahn)	Sacher (2016)
11	Healthy control group	70		Schmid (2012); Schwab (2016); Sedlaczek (2016)
BoDyS.IT	Diverse (Italian speaking)	40	Dipartimento di Scienze Chirurgiche, Università di Torino, Italy (P. Cancialosi)	Protti (2014)

Note. Most of the references cited are student theses.