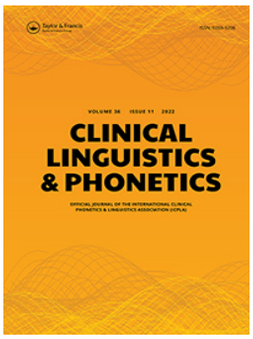

Clinical Linguistics & Phonetics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/iclp20

Clinical measures of communication limitations in dysarthria assessed through crowdsourcing: specificity, sensitivity, and retest-reliability

Katharina Lehner, Wolfram Ziegler & for the KommPaS Study Group

To cite this article: Katharina Lehner, Wolfram Ziegler & for the KommPaS Study Group (2022) Clinical measures of communication limitations in dysarthria assessed through crowdsourcing: specificity, sensitivity, and retest-reliability, Clinical Linguistics & Phonetics, 36:11, 988-1009, DOI: 10.1080/02699206.2021.1979658

To link to this article: https://doi.org/10.1080/02699206.2021.1979658

Published online: 12 Nov 2021.

CLINICAL LINGUISTICS & PHONETICS
2022, VOL. 36, NO. 11, 988–1009
https://doi.org/10.1080/02699206.2021.1979658

Taylor & Francis
Taylor & Francis Group

Clinical measures of communication limitations in dysarthria assessed through crowdsourcing: specificity, sensitivity, and retest-reliability

Katharina Lehner, Wolfram Ziegler for the KommPaS Study Group

Clinical Neuropsychology Research Group, Institute of Phonetics and Speech Processing, Ludwig-Maximilians-University, Munich, Germany

ABSTRACT

Assessing the impact of dysarthria on a patient's ability to communicate should be an integral part of patient management. However, due to the high demands on reliable quantification of communication limitations, hardly any formal clinical tests with approved psychometric properties have been developed so far. This study investigates a web-based assessment of communication impairment in dysarthria, named *KommPaS*. The test comprises measures of intelligibility, naturalness, perceived listener effort and communication efficiency, as well as a total score that integrates these parameters. The approach is characterized by a quasi-random access to a large inventory of test materials and to a large group of naïve listeners, recruited via crowdsourcing. As part of a larger research program to establish the clinical applicability of this new approach, the present paper focuses on two psychometric issues, namely specificity and sensitivity (study 1) and retest-reliability (study 2). Study 1: *KommPaS* was administered to 54 healthy adults and 100 adult persons with dysarthria (PWD). Non-parametric criterion-based norms (specificity: 0.95) were used to derive a standard metric for each of the four component variables, and corresponding sensitivity values for the presence of dysarthria were identified. Overall classification accuracy of the total score was determined using a ROC analysis. The resulting cutscores showed a high accuracy in the separation of PWD from healthy speakers for the naturalness and the total score. Study 2: A sub-group of 20 PWD enrolled in study 1 were administered a second *KommPaS* examination. ICC analyses revealed good to excellent retest reliabilities for all parameters.

ARTICLE HISTORY
Received 8 March 2021
Revised 5 August 2021
Accepted 5 September 2021

KEYWORDS
Dysarthria; communication limitation; web-based assessment; crowdsourcing; psychometric evaluation

Introduction

The gold-standard of clinical dysarthria diagnostics is based on auditory-perceptual assessments addressing the characteristics of impaired respiration, phonation, articulation, resonance, and prosody, as applied in common assessment tools such as the *Frenchay Dysarthria Assessment* (FDA-2; Enderby & Palmer, 2011) or the *Bogenhausen Dysarthria Scales* (BoDyS; Ziegler et al., 2017), or in less formal examination protocols, as proposed by Kent (2009) or Duffy (2020). Beyond this, lasting efforts are made to supplement auditory-perceptual methods by clinically usable acoustic analyses (e.g., Laganaro et al., 2021). These

CONTACT Katharina Lehner katharina.lehner@ekn-muenchen.de; wolfram.ziegler@ekn-muenchen.de Institut Für Phonetik Und Sprachverarbeitung, LMU, Schellingstraße 3, München 80799, Germany

diagnostic approaches have in common that they rely on function-oriented, analytically guided examinations and require professional experience in motor speech disorders. They are indispensable in the clinical care of patients with dysarthria, as they provide guidance for tailored interventions that target the suspected underlying dysfunction.

Yet, therapeutic care must also consider the limitations that are most consequential to persons with dysarthria (PWD) in everyday life, namely not being understood, being stigmatized for their conspicuous manner of speaking, and being shunned because conversing with them is exhausting and time-consuming (e.g., Dykstra et al., 2007; Schölderle et al., 2019; Walshe & Miller, 2011). After all, the overall goal of treatment is to reduce these limitations and thereby improve the patients' social participation and quality of life (e.g., Gurevich & Scamihorn, 2017). Dedicated assessments are needed to determine whether and to what extent the speech symptoms detected by the expert's trained ear have an impact on communication abilities, since the recognition and assessment of these symptoms alone does not allow a straightforward prediction of their communicative relevance (e.g., Schölderle et al., 2016). Moreover, measuring the impact of dysarthria on communication helps to ensure that functionally motivated therapeutic adjustments are effective at an ecologically relevant scale.

In clinical or private practice settings, it is virtually impossible to assess communication impairments in all their facets, i.e., how PWD experience them in real life (Kent & Kim, 2010). Clinicians may resort to self-report questionnaires to inquire in detail a patient's dysarthria-related communication and participation limitations (e.g., Walshe et al., 2009), but such instruments are usually not applicable in the acute stage of dysarthric impairment, nor are they flexible and sensitive enough to follow in short time intervals the natural course or recovery of dysarthric impairment or the effectiveness of treatment. Therefore, a tool is needed that is more readily usable to assess the impact of a patient's disorder on communication-relevant speech parameters. Such an instrument should be widely accessible in the health care system, feasible at the therapist's desk, and applicable to individuals with all types and severities of dysarthria at all stages of their disease. Moreover, as a diagnostic tool it must provide replicable, reliable, and valid results.

The study presented here deals with psychometric properties of a web-based diagnostic instrument named *KommPaS* (German acronym for **Comm**unication related **Pa**rameters in **S**peech Disorders; Lehner & Ziegler, 2021) that was developed to assess communication relevant parameters of dysarthria in a clinical standard setting. The assessment comprises three auditory perceptual parameters that have played a dominant role in the literature, i.e., *intelligibility, naturalness*, and *perceived listener effort*. Moreover, as an extension, *communication efficiency* is assessed as a parameter that relates intelligibility to the time needed to convey an utterance. Finally, appropriately standardized scores of these parameters are aggregated to a *communication total score* that integrates across a patient's test profile and serves as an overall communication index within the *KommPaS* framework.

A key feature of *KommPaS* is that it is based on human listeners rather than on acoustic measures. In the past, acoustic markers of intelligibility have been examined using numerous parameters such as formant measures of vowel articulation (Kent et al., 1989; Lansford & Liss, 2014a), spectral measures (Janbakhshi et al., 2019), or speech rhythm metrics (Selouani et al., 2012; for a review see Feenaughty et al., 2014). Despite the enormous progress of these efforts, reliance on human listeners in clinical assessment is still indispensable, because there

is no compelling evidence that acoustically based diagnostic instruments can mimic human perception of communication-related dimensions of dysarthric speech to a sufficiently reliable and valid degree across all types and severities of dysarthria.

Another general feature of *KommPaS* is that it provides continuous metrics for each of its component parameters rather than a classification based on conventional multivariate statistics (e.g., Lansford & Liss, 2014b) or machine learning algorithms (e.g., Tsanas et al., 2012). A major reason for this is that the mere classification of a speaker as dysarthric or non-dysarthric is of secondary importance for clinical purposes. It is more important to derive criterion-based metrics that describe the severity of impairment on each of the four dimensions of the *KommPaS* profile on commensurable scales, and to obtain a score that integrates these metrics into an overall measure of communication impairment.

Among the parameters included in the *KommPaS* profile, the most widely discussed index of communication impairment in dysarthria is intelligibility, understood as *the extent to which a speaker's acoustic signal can be accurately recovered by a listener* (c.f. Hustad, 2008). Since the primary goal of communicative interaction is to make oneself understood, any reduction of intelligibility due to dysarthria implies a major limitation in everyday life. Yet, dysarthria cannot be equated with loss of intelligibility, because intelligibility is not necessarily compromised in all speakers with dysarthria (e.g., Laganaro et al., 2021; Schölderle et al., 2016), and because communication impairment in dysarthria includes more than just the potential limitation of intelligibility. Especially mildly impaired PWD are difficult to distinguish from healthy speakers only on the basis of intelligibility (Yorkston & Beukelman, 1981).

In research, three other traits were proposed as indicators of a communication impairment in dysarthria. First to mention is *naturalness* as a broadly accepted parameter. It describes *to what extent the way someone speaks sounds natural and not irritating, i.e., whether it conforms to the expected standard of unimpaired speech* (c.f. Yorkston, 2010) (see Klopfenstein et al., 2020 for a systematic overview). Since a person's individual way of speaking is part of their identity and personality (Scherer, 1972), any dysarthric impairment will inevitably affect the patient's individual "acoustic fingerprint" and, in severe cases, result in an irritating or even bizarre, stigmatizing manner of speaking, with far-reaching social consequences (Schölderle et al., 2019).

Two further parameters strongly associated with intelligibility are *perceived listener effort*, i.e., *the perceived amount of effort expended to understand the speaker* (c.f. Whitehill & Wong, 2006), and *communication efficiency*, expressed by the *ratio between the number of intelligible syllables or words and utterance duration* (c.f., Yorkston & Beukelman, 1981). Whereas conversational interactions among non-impaired speakers are mostly effortless and quick, communicating with a person with dysarthria can be challenging and tiring when the conversation partner must concentrate on decoding the patient's utterances and tolerate their slow and dysfluent speaking. Therefore, both an increase of *perceived listener effort* and a decrease of *communication efficiency* can cause frustration on the interlocutor's side and a tendency to avoid conversations with patients who are dysarthric.

There is clear evidence in the literature that, in addition to intelligibility, naturalness (e.g., Dagenais et al., 2006; Schölderle et al., 2016), perceived listener effort (Beukelman et al., 2011, 2014; Whitehill & Wong, 2006) and communication efficiency (Yorkston & Beukelman,

1981) have their own clinical relevance in dysarthria assessment. However, only the latter has ever been considered in a clinical test protocol (c.f., SIT; Yorkston et al., 1996),[1] whereas the first two have not been included in any standard clinical dysarthria test so far.

Methodological standards for a reliable and valid measurement of intelligibility and other communication-relevant parameters are high, as there are many factors, beyond the quality of the speech output itself, that must be controlled because they may enhance or impede a listener's perception of dysarthric speech. Two factors are particularly relevant and are therefore in the centre of the test construction that is validated in this article.

One concerns the *selection of listeners* to be consulted for the evaluation of communication relevant parameters. It is known that a listener's degree of familiarity with the speaker (DePaul & Kent, 2000) and their perceptual experience with dysarthria in general (Borrie et al., 2017; Dagenais et al., 1999; Smith et al., 2019) have a strong impact on how dysarthric speech is perceived and assessed. These findings disqualify the attending speech and language therapists (SLTs) for the clinical assessment of communication-related parameters for two reasons: First, as experts, they are adapted to the characteristics of dysarthric speech in general, which curtails the representativeness of their judgments. Second, because of their familiarity with the individual PWD to be diagnosed, they are particularly adapted to the PWD's speech and become increasingly biased over time.

The second factor that influences the perception of dysarthric speech, especially its intelligibility, are the *materials* used in a test, e.g., the degree of a listener's foreknowledge of the utterances to be judged (e.g., Utianski et al., 2011), the predictability through syntactic and semantic context (e.g., Beverly et al., 2010), or word-level factors like lexical frequency, lexical familiarity, neighborhood density or articulatory complexity (e.g., Lehner & Ziegler, 2021). To avoid material-based biases, test materials, i.e., word or sentence lists, must be systematically controlled for these factors, but at the same time also be highly diversified in order to create maximum uncertainty and avoid learning effects on the part of the test listeners.

Although listener- and material-related factors have been examined extensively in experimental studies, especially on intelligibility (Barreto & Ortiz, 2020; Beverly et al., 2010; Chiu & Forrest, 2018; McAuliffe et al., 2017, 2013; Smith et al., 2019; Tjaden & Wilding, 2011; Utianski et al., 2011) none of the established clinical assessment procedures controls for all these factors to a satisfactory degree. Typically, intelligibility is assessed by a speech-language pathologist, usually the therapist responsible for the patient being assessed (e.g., MonPaGe, Laganaro et al., 2021; SIT, Yorkston et al., 1996) which can lead to the adaptation bias mentioned above. Moreover, the number of the speech materials used for the assessment is often limited, which makes the stimuli predictable (e.g., FDA-2; Enderby, 2011), and the test words or sentences are hardly ever controlled for influential variables. This is especially the case in informal procedures, e.g., judgments of the intelligibility of spontaneous speech or a standard reading passage, which are common in clinical practice (c.f., Gurevich & Scamihorn, 2017).

The *KommPaS* approach takes both of these problems into account: First, regarding the recruitment of test listeners, an online crowdsourcing approach is taken to draw non-expert, naïve informants from a large pool of crowdworkers. This makes it possible to reach,

[1]More recently, Vojtech et al. (2019) examined communication efficiency in an experimental study of synthesized speech that might inform future developments of augmentative and alternative communication devices.

within a short time, a substantial number of listeners with whom a person with dysarthria is likely to interact in everyday life and whose non-professional 'real-life' perspective renders their judgment ecologically valid (McAllister Byun et al., 2015; Nightingale et al., 2020). Unlike conventional experimental methods of involving naïve listeners by recruiting a fixed panel of laboratory volunteers, crowdsourcing offers a scalable and viable solution to engage laypersons in clinical assessment from the speech therapist's desk. Second, online access to a labelled database comprising more than 12,000 content words and more than 500 syntactically and semantically neutral carrier phrases permits the ad-hoc compilation of highly diverse, but systematically designed sentence lists for each test administration. Thereby, effects of potential foreknowledge or predictability of the test materials and of linguistic features like utterance length or lexical frequency are nullified (Lehner & Ziegler, 2021). The assessment procedure will be described step-by-step in the methods section.

The above-mentioned diversification of test materials and test listeners, which distinguishes the *KommPaS* approach from earlier clinical assessments, raises the question if the completely automated, quasi-random selection of test words, carrier phrases, and listener panels can be reconciled with the psychometric standards required in clinical diagnostics. In particular, it must be ruled out that the employed 'controlled diversification' of listeners and materials has an adverse effect on the discriminatory potential of the test and its retest-reliability. In this article, two studies are reported that address these two major psychometric issues.

The objectives of **Study 1** were

(1) to establish cutoff values ensuring a reasonable specificity of the four component parameters *intelligibility, naturalness, perceived listener effort*, and *communication efficiency* and to determine the rates of PWD who perform within a normal range on each of these dimensions,
(2) to derive standardized metrics measuring a speaker's distance from neurotypical control speakers on each dimension, and
(3) to determine whether the *KommPaS* total score, as a measure that combines the four dimensions, is accurate enough to yield a reasonable distinction between dysarthric and unimpaired speech.

As regards to **objective (1)**, a biased emphasis on specificity over sensitivity was chosen, because the motor impairments underlying dysarthria do not necessarily have adverse consequences for each of the four *KommPaS* dimensions individually. As an example, it is known that by far not every PWD has an intelligibility problem (e.g., Dagenais et al., 2006; Schölderle et al., 2016). Following conventions of criterion-based testing, a rate of false positives of at most 5% in a sample of neurotypical control speakers was operationally defined as cutoff. These cutoffs were then used to determine the proportions of PWD that fall within the normal range on each dimension (sensitivity). The following hypotheses were made:

(a) Regarding *intelligibility*, *perceived listener effort* and *communication efficiency*, a considerable proportion of PWD will score within the 95% range of neurologically healthy speakers, because dysarthria does not always lead to problems in each of these domains. That is, given a pre-defined specificity of 0.95, these parameters will only have a moderate sensitivity.

(b) Regarding speech *naturalness*, only a much smaller proportion of the patients will score within the 95% normal range, because naturalness is expected to be affected by even mild impairments of any of the functional components of speech, i.e., respiration, voice, articulation, resonance, and prosody, and is therefore highly vulnerable to dysarthric impairment. That is, given a pre-defined specificity of 0.95, naturalness is expected to have the highest sensitivity among the four *KommPaS* variables.

As regards to **objective (2)**, the distributions of the four *KommPaS* parameters in neurologically healthy controls were used to calculate non-parametrically standardized variables measuring an individual's distance from neurotypical control speakers. These transformations made the four parameters comparable and allowed them to be combined into a *communication total score* representing a person's overall performance across the four *KommPaS* dimensions.

Regarding **objective (3)**, i.e., to determine the accuracy of the *total score*, a different approach was taken. In this case, a plausible a-priori assumption was that patients with dysarthria are likely to be impaired on at least one of the four *KommPaS* parameters, suggesting that the *communication total score* should separate PWD from healthy speakers with a high accuracy. Therefore, the predetermination of a 0.95-specificity value was skipped and a ROC analysis was performed to balance sensitivity and specificity simultaneously. On this account, the following hypothesis was made:

(c) Empirically determined specificity and sensitivity values for the *communication total score* yield a clinically meaningful separation of speakers with dysarthria from neurologically unremarkable speakers and render the KommPaS total score a parameter with a high classification accuracy.

The objective of **Study 2** was to examine whether the pseudo-random selections of test listeners and test materials, that are major principles of the *KommPaS* test design, would not spoil the reliability of the parameters. More specifically, we asked whether two tests of the same speaker under the same conditions, but with different materials and different listeners, yield the same results for *intelligibility, naturalness, perceived listener effort, communication efficiency* and the *communication total score* (test-retest reliability).

Study 1: specificity, sensitivity and standard metrics

Methods

Participants

Persons with dysarthria (PWD). In total, the study involved 100 persons with dysarthria (39 women, 61 men; 53.3 ± 18.1 years) who were recruited from twelve different rehabilitation clinics in Germany (also reported in Lehner & Ziegler, 2021). The inclusion criteria were: (1) age 18 to 80 years, (2) German as first language, (3) no structural damage of the respiratory, laryngeal or articulatory organs, (4) no associated aphasia or apraxia of speech, and (5) no manifest dementia that might prevent them from performing the task. There were no limitations regarding the etiology or severity of the dysarthria.

Table 1. Demographic data and severity of dysarthria of the patient sample, by etiologies.

Etiology	N	Age mean (sd); range in years	Gender F/M	Severity of dysarthria (profile height of the *BoDyS*)[a] mean (sd); range
Stroke	25	60.1 (16.6); 18–79	11/14	44.6 (11.6); 21.0–62.0
Cerebral palsy	18	35.4 (12.1); 18–64	12/6	39.7 (7.5); 26.0–51.0
Parkinson's Disease	14	66.9 (10.6); 46–80	5/9	53.7 (9.5); 32.0–65.0
Atypical Parkinson's syndromes	14	67.3 (9.8); 49–80	4/10	48.0 (9.1); 25.0–60.0
Traumatic brain injury	11	38.9 (19.1); 18–68	2/9	41.6 (9.5); 26.0–58.0
Multiple sclerosis	9	49.9 (6.2); 40–59	2/7	52.9 (6.8); 43.0–63.0
Other	9	48.9 (16.5); 21–72	3/6	38.9 (7.0); 31.0–53.0
Total	100	53.4 (18.1); 18–80	39/61	45.0 (10.4); 21.0–65.0

[a]based on standard T-norms (mean 50 ± 10) from a calibration sample of 220 PWD; Ziegler et al. (2018). A T-score of 50 represents the mean profile height of the calibration sample (see Crocker & Algina, 2008; Chapt. 19).

Table 1 gives an overview of the patients' demographic data and overall severity of dysarthria, broken down by etiologies. As a measure of overall severity, the profile elevation of the standard-normalized *Bogenhausen Dysarthria Scales* (*BoDyS*) was used (Ziegler et al., 2018, 2017).[2]

Neurologically healthy control speakers (CON). The study included a control group of 54 neurologically healthy speakers with unimpaired speech. The group was stratified by age and gender, i.e., nine females and nine males each in three age groups from 20–40, 41–60 and 61–80 years (mean age: 48.3 ± 16.5 years). The criteria for participation were: (1) German as first language, (2) no neurological or neuropsychological disease, (3) no articulation, voice, or fluency impairment.

Test procedure – KommPaS WebApp

Examination setting. Persons with dysarthria were tested online in a quiet room in their respective clinical institutions by the therapist in charge, or in some cases by the first author. A Sennheiser SC 40 USB monaural headset microphone and a laptop with Wifi connection were used as equipment. Testing of the healthy control group was carried out by the first author under the same conditions.

Stimuli. As stipulated in the *KommPaS* protocol, each speaker session comprised 30 sentences in total. It started with always the same short natural text consisting of three sentences,[3] which later served for the speech naturalness ratings. Thereafter, 27 sentences (target words embedded in carrier-phrases) were presented which were automatically selected from a large labeled database comprising more than 12,000 one- to three-syllabic German content words and a set of more than 500 syntactically and semantically neutral carrier phrases of 4 to 6 syllables length (e.g., ' – – is easy to understand'; for examples see Table 4). The selection algorithm ensured that for each examination a new sample of test

[2]The *BoDyS* evaluations for the patient sample were completed by three staff members (among them the first author) with specific training in the administration of the *BoDyS*. The reliability of the *BoDyS* scales was approved in Ziegler et al. (2017).

[3]'Der Englische Garten ist einer der größten Parks der Welt. Er liegt in München, der Hauptstadt von Bayern. Im Park gibt es lange Spazierwege und mehrere Biergärten.' [The English Garden is one of the largest parks in the world. It is located in Munich, the capital of Bavaria. In the park there are long walks and several beer gardens]

sentences was composed, balanced for word length, word frequency, sentence length, and embedding position (initial, medial, final). Participants were asked to read aloud or repeat the test stimuli that appeared one by one on the screen after starting the examination.

Evaluation by crowdworkers. Immediately after the examination was finished, the speech samples of each *KommPaS* session were automatically integrated into a listening task template. Under strict considerations of data protection, a remotely working moderator of the web app (first author) placed the web-link to the task on the commercial microtasking platform *Clickworker*, where the registered crowdworkers could accept the offer asynchronously.

Crowdworkers were to meet the following criteria for participation: (a) German native speakers, (b) between 18 and 60 years of age, (c) no hearing impairment (self-reported), and (d) no or only incidental experience in communicating with individuals with neurogenic speech disorders (self-reported). A total of 466 different crowd listeners[4] took part in study 1. Crowdworkers received 1.75 € per completed listening task. With a median task duration of 9 minutes, the payment was above the German minimal wage.

The crowdworkers were first presented the short introductory 3-sentence-text spoken by a patient to estimate the *naturalness* of the patient's speech on a visual analog scale (VAS; horizontal bar with the anchors "very unnatural" (left, =0) and "natural" (right, =100)). There was a written instruction that paraphrased the description of the naturalness criterion given above. Furthermore, six audio samples providing an impression of how naturalness can vary in PWD were presented. This was followed by a word transcription task for the assessment of *intelligibility*. The crowdworkers listened to each of the 27 sentences once and in the same order as they were spoken by the patient. After each sentence audio and a short pause, the written carrier sentence appeared on the screen (e.g., '_____ is easy to understand.'), and listeners were instructed to type the target word in the respective cloze gap (e.g., '*Summer* is easy to understand.'). Prior to this, crowdworkers had to complete a sample task, i.e., process one sentence spoken by a healthy model speaker, to familiarize themselves with how the transcription task works. There was no training or familiarization with dysarthric speech otherwise. The proportion of correctly transcribed words was recorded as the listener's individual intelligibility score for the speaker. After the 27 test items were completed, listeners were again presented a VAS to estimate the *perceived effort* they had expended to understand the speaker.

Each speaker's test was evaluated by a panel of nine crowdworkers, which is the recommended listener number for crowd-based perceptual studies (McAllister Byun et al., 2015).

Aggregation of crowdworker scores. Since listeners recruited by crowdsourcing may vary widely in how reliably they complete a task, an algorithm was developed to weight listeners for their relative performance on the transcription task. For each panel of nine listeners, the crowdworker who attains the best (i.e., highest) score is assumed to be the most 'faithful' listener, and the corresponding score is assigned a weight of 1. All other listeners are

[4]They represent a subgroup of over 250,000 German speaking crowdworkers registered at *Clickworker*. *Clickworker* states that their community is a broad cross-section of the population in terms of age or employment level. Furthermore, the distribution of gender with 51% men and 49% women is roughly balanced (Clickworker, 2021).

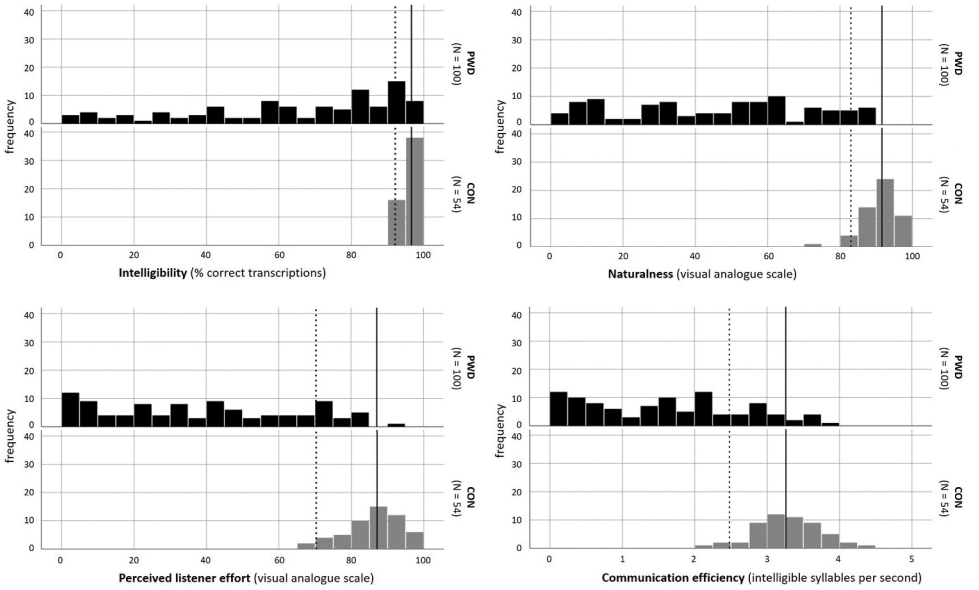

Figure 1. Distributions of the *KommPaS*-scores for *intelligibility, naturalness, perceived listener effort* and *communication efficiency* in persons with dysarthria (PWD; black) and in healthy control speakers (CON; light grey). Vertical solid and dashed lines: median values and 5th percentiles, respectively, of the CON group.

weighted by a function of their distance from the best listener, more specifically, a negative exponential of the distance. By this algorithm, the scores of spammers or extremely inattentive listeners are given exponentially low weights. This aggregation method has proved to be insensitive to cheating and outperformed other, conventional aggregation methods (including median and arithmetic mean) in terms of accuracy and stability across large numbers of different listener panels assessing the same speech samples (Ziegler et al., 2021). The listener weights determined from the intelligibility scores were considered as a general index of listener faithfulness and were therefore also applied in the aggregation of the nine individual scores of *naturalness* and *perceived listener effort*.

KommPaS profile and standard metrics. To complete the KommPaS profile, the weighted averages of *intelligibility, naturalness*, and *perceived listener effort* are supplemented by a fourth, derived variable, i.e., *communication efficiency*. *Communication efficiency* (EFF), with the dimension *intelligible syllables per second*, is determined as

$$\mathrm{EFF} = \mathrm{RATE} * \mathrm{INT}/100, \tag{1}$$

with RATE denoting an individual's speech rate (syllables per second) averaged across the 27 sentences of the intelligibility task,[5] and INT denoting intelligibility in %. For instance, a speaker with an *intelligibility* score of 85% and a speech rate of 1.5 [syll/sec] obtains a *communication efficiency* score of 0.85 * 1.5 [syll/s] = 1.28 [intellsyll/s].

[5]Speech rate is measured automatically during the upload of the recording by detecting the start and end positions of the spoken sentences using the software ffmpeg and calculating the individual durations of the speech samples. The start and end positions are manually reviewed and corrected in case of errors.

Table 2. Influence of speaker age on the four component variables of KommPaS in 54 neurotypical adult speakers between 19 and 77 years. n.s.: not significant (α = 0.05); ** $p < .01$; *** $p < .001$.

Variable	$R2_{adj}$	F(1,52)	Slope
Intelligibility	0.01	1.5 n.s.	0.000
Naturalness	0.16	11.3 ***	−0.001
Perceived listener effort	0.04	2.9 n.s.	−0.001
Communication efficiency	0.14	9.8 **	−0.012

To make the four component parameters of the KommPaS profile comparable, they were standardized relative to their respective distributions in the control group of neurotypical speakers. Non-parametric transformations were performed by calculating the distance of each participant's individual score from the median value of the healthy speakers, in units determined by the dispersion between the median and the 5th percentile (i.e., the distance between the two vertical lines in the four panels of Figure 1). More formally, for each *KommPaS* variable v_i (i = 1,.,4) a standardized variable v_i' was determined as

$$v_i' = (m_i - v_i) / \left(m_i - p_i^{5\%}\right), \tag{2}$$

where m_i denotes the median value and $p_i^{5\%}$ the 5th percentile of the variable v_i in the control group. Similar to a z-transform for parametric data, these transformations make the four variables commensurable. Note that standardized scores above 1.0 indicate performance outside the 95% range of neurotypical speakers, with larger scores indicating more severe impairment.[6]

Communication total score. Finally, a *communication total score* was determined as an overall measure of communication impairment by averaging the four standardized *KommPaS* variables.

Results

In order to establish criterion-oriented norms, the control participants' raw scores of *intelligibility, naturalness, perceived listener effort* and *communication efficiency* were submitted to linear regression analyses to determine potential influences of age. Linear regression models were fitted using the function 'lm' of the R-package *stats* (RCoreTeam, 2020). The results revealed no age effect for *intelligibility* and *perceived listener effort*, and small but significant effects for *naturalness* and *communication efficiency* (Table 2). Since there was no effect of gender on any of the four variables (two-sided t-tests, $|t| < 1.6$, $p > .05$ in all cases), all norms were gender-independent.

In order to compensate for these effects, the raw scores of *naturalness* and *communication efficiency* were normalized for age, with a reference age of 50 years and a bonus/malus of 0.1% per year for *naturalness* and 0.012 [syll/s] per year for *communication efficiency*, respectively,

[6]In this article, only raw scores are reported for the four component variables of the *KommPaS* profile. Standardized scores are only used for the computation of the total score.

depending on a speaker's age distance from 50 (Table 2, rightmost column). In the following, only age-transformed raw data will be reported for *naturalness* and *communication efficiency*, while *intelligibility* and *perceived listener effort* will remain untransformed.

Figure 1 depicts the distributions of these scores for the individuals with dysarthria (PWD, n = 100, top panels) and the control group (CON, n = 54, bottom panels). The vertical lines denote the median values (solid lines) and the 5th percentiles (dashed lines), respectively, of the control group.

The patients' *intelligibility* scores covered the entire range from 0 to 98% (median 72%), with a fairly left-skewed distribution (skewness: −0.69). Remarkably, the healthy participants were not all 100% intelligible (range 92% – 100%, median 97%; 5th percentile 92%). Sixteen patients scored above the 5% cutoff value, amounting to a sensitivity of only 0.84 of the intelligibility score in distinguishing between the CON and the PWD groups.

Regarding *naturalness*, the healthy participants ranged between 74 and 97 (median 92; 5th percentile: 83), while the speakers with dysarthria received rather uniformly distributed values between 0 and 90 (median 48). Only 7/100 patients had naturalness scores above the CON group's 5% threshold, corresponding to a sensitivity of 0.93 of the *naturalness* score to distinguish between the two groups. Overall, the rate of correctly classified participants as dysarthric or non-dysarthric was 145/154 = 0.94.

With regard to the variable *perceived listener effort*, it is notable that understanding the control participants was by far not always perceived as completely effortless. The scores for the CON group ranged between 67 and 98 (median 87), with a 5% cutscore at 70. The PWD group received values between 8 and 95 (median 37). Among them, 17 patients were judged better than the cutscore, corresponding to a sensitivity of 0.83 of *perceived listener effort* for the presence of dysarthria. Unlike the *intelligibility* scores, the *perceived listener effort* scores had a slightly right-skewed distribution (skewness: 0.20).

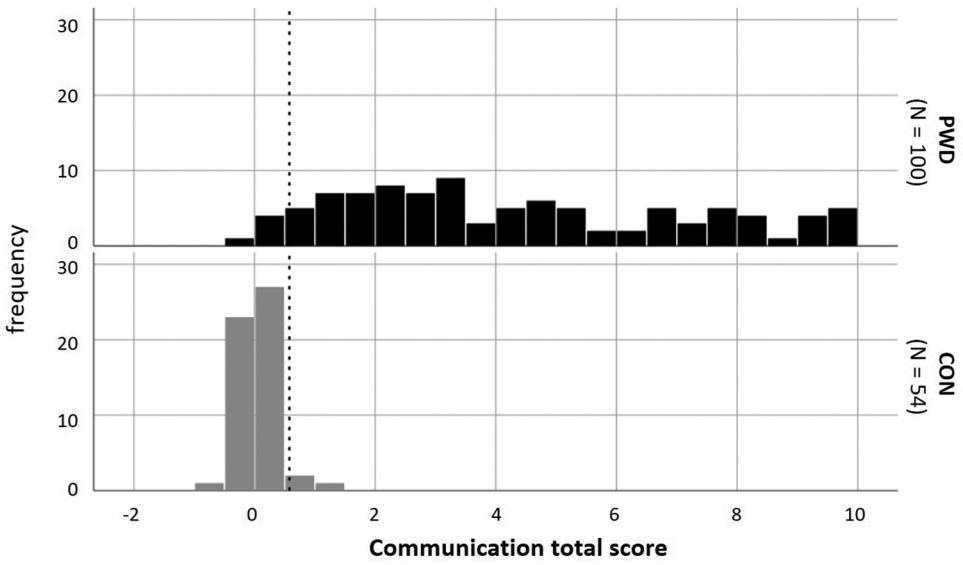

Figure 2. Distribution of the *communication total score* for persons with dysarthria (top) and healthy control speakers (bottom). Vertical dashed line: cutscore obtained by a ROC-analysis.

Table 3. Distributions of KommPaS scores of the control participants (CON) for each of the five *KommPaS* parameters and corresponding sensitivity and specificity coefficients. Note that specificity was deliberately set to 0.95 for the four component parameters (lines 1 to 4) and determined through a ROC analysis for the total score (line 5).

Parameter	Median CON[1] (n = 54)	Cutscore		Sensitivity/ specificity
Intelligibility	97	5th percentile	92	0.84/0.95
Naturalness	92	5th percentile	83	0.93/0.95
Perceived listener effort	87	5th percentile	70	0.83/0.95
Communication efficiency	3.26	5th percentile	2.48	0.77/0.95
Communication total score	0.02	ROC[2] threshold (Youden's J)	0.58 0.98	0.96/0.95
		AUC[3]	[0.97–1.00]	

1 control group
2 receiver operating characteristics
3 area under curve [95% CI]

The distributions of the *communication efficiency* scores had the largest overlap. The PWD group showed a rather flat, again slightly right-skewed distribution between values of 0 and 3.82 (median 1.69; skewness: 0.12), while the healthy participants ranged between 2.15 and 4.31 (median 3.26), with a 5% cutoff at 2.48. Twenty-three patients with dysarthria had *communication efficiency* scores above the cutscore, corresponding to a sensitivity of only 0.77.

Figure 2 shows the distribution of the *communication total score* for the CON and the PWD group separately.

Recall that this parameter measures the average distance from the healthy speakers' median values across the four *KommPaS* variables, as defined in equation (2) above. Therefore, higher values indicate more severe impairment.

The speakers with dysarthria covered a wide range of *communication total scores*, from 0.12 to 9.95 (median 3.67). The participants of the CON group were distributed close to zero with a median value of 0.01 (range: −0.47 to 1.14).

Unlike in the four single variables reported above, where we aimed to determine cutscores ensuring specificity values of 0.95, it is reasonable for the *communication total score* to seek an optimal balance between sensitivity and specificity (see hypothesis (c) in the Introduction). A ROC analysis[7] with an excellent AUC-value of 0.98 (95% CI: 0.97– 1.00) revealed a cut-off of 0.58 (using Youden's J as a criterion), resulting in a sensitivity of 0.96 and a specificity of 0.95. Five PWD reached false-negative scores, while two healthy speakers received false-positive scores. With a total of 7/154 misclassifications, the *communication total score* reached a known-groups validity of 0.96 in separating PWD from the healthy participants included here.

Table 3 summarizes the cut-off values for each of the five parameters.

Discussion

Regarding the distributions plotted in Figure 1 it is notable that the healthy participants were not always 100% intelligible. This result is not unexpected as it is in line with previous descriptions (e.g., Miller, 2013; Stipancic et al., 2016). It is also plausible considering the type of task used in this study, i.e., word recognition in the absence of any context, which is known to cause sporadic errors even in understanding undisturbed speech (e.g., MacDonald & Hsiao, 2018). At the same time, several PWD had comparably high

[7]The ROC analysis was performed using the R-package 'pROC' (Robin et al., 2011).

intelligibility scores, which is also not surprising against the backdrop that not every person with dysarthria has an intelligibility problem. As hypothesized in the introduction, this resulted in only a moderate sensitivity when a cutscore was determined to secure a specificity of 0.95.

As was also predicted, a similarly moderate sensitivity was obtained for the *perceived listener effort*-score. Compared with the *intelligibility* scores, the *perceived listener effort* ratings were inversely skewed, suggesting that low levels of *perceived listener effort* were considerably more frequent than low intelligibility levels. As a careful interpretation, this gives the two parameters a complimentary clinical relevance (c.f., Miller, 2013). Some speakers were eventually able to reach decent intelligibility scores, but only at the expense of a strong cognitive effort on the part of the listeners. This finding coincides with the results of other studies comparing intelligibility with listener effort (e.g., Beukelman et al., 2011, 2014).

Communication efficiency reached the lowest sensitivity value among the four *KommPaS* parameters. Considering the two components contributing to *communication efficiency*, i.e., intelligibility and speaking rate, the large overlap of the PWD and CON groups was ascribable to the fact that the healthy participants had rather low speaking rates. Speaking rates in typical speakers are known to vary with the linguistic task (Yorkston, 2010), and the elicitation paradigm used in *KommPaS*, i.e., sentence repetition or reading, obviously provoked a very measured tempo in many of the healthy participants. Yet, the comparably large proportion of patients with considerably low *communication efficiency* scores suggests that the parameter nonetheless provides additional clinically relevant information in patients with severe impairment who are still reasonably intelligible, but extremely inefficient due to a concomitant low speaking rate.

In accordance with the hypotheses formulated in the introduction, the highest sensitivity was obtained for the *naturalness* score. This result is highly plausible, since even mild impairments of any of the motor subsystems of speech, i.e., respiratory, laryngeal, velopharyngeal and articulatory, as well as any prosodic aberration, signals a deviation from natural sounding speech. The finding that 9/100 subjects with dysarthria were fully intelligible but perceived as unnatural gives the *naturalness* scale independent clinical relevance (c.f. Dagenais et al., 2006; Schölderle et al., 2016). The *naturalness* scores of the healthy persons were more distributed than the *intelligibility* scores, presumably because naturalness ratings leave room for more sensible assessments of speaker idiosyncrasies, physiological variants, or regional dialects which are within the range of speech features in the neurologically healthy population and therefore diagnostically irrelevant (Klopfenstein et al., 2020).

Overall, and as hypothesized in the introduction, the four communication-related parameters of the *KommPaS* profile turned out to have different sensitivities when cutscores securing a specificity of 0.95 were determined. This result reflects that persons with dysarthria are not necessarily restricted in all aspects of speech communication to the same extent. However, when the four scales were combined to form a common score of overall communication impairment, a high accuracy of separating healthy speakers from speakers with dysarthria was achieved.

Study 2: Test-Retest-Reliability

Methods

Participants

To examine the retest reliability of the *KommPaS* variables, a convenience sample of 20 patients selected from the larger group in Study 1 underwent a second KommPaS examination immediately after the first examination reported in Study 1. The decision to select a patient for a second testing was made ad-hoc during the first assessment, considering patient-related (e.g., obvious fatigue) and organizational criteria (e.g., willingness to participate again, availability in time). Moreover, care was taken over time that the patients selected for a retest covered a sufficiently broad range of dysarthria severity, according to the clinical impression obtained during the first assessment. Age, gender, and etiology were not taken into consideration.

The sample ultimately included 6 women and 14 men, aged 53.9 ±10.8 and encompassed patients with various neurological diseases, i.e., stroke (n = 6), cerebral palsy (n = 1), Parkinson's Disease (n = 1), atypical Parkinson's syndromes (n = 4), traumatic brain injury (n = 1), multiple sclerosis (n = 5) and other (n = 2), and therefore represented all etiologies that were also covered in the main sample.

Procedure

The two examinations were performed in immediate succession with only a short pause in between, in order to exclude drug-induced variations (e.g., on-off patients with PD), therapy-related effects, daily fluctuations or variations due to disease progression or relapsing-remitting episodes (e.g., in multiple sclerosis). Recording equipment and environmental conditions were the same as in study 1. As specified in the KommPaS intelligibility protocol, the two examinations were based on different stimulus sets randomly selected from the *KommPaS* word and sentence database according to the rules specified in the methods section of Study 1 (see Table 4 for an

Table 4. Excerpts from two sample lists of test items. For each test application, a new list of 27 items is generated. Lists are compiled through random selections from repositories of 540 carrier phrases and more than 12,000 German 1- to 3-syllabic words, labelled for word length and lexical frequency (subtlex-np; Lehner & Ziegler, 2021). For each row in the table (i.e., each combination of word length and lexical frequency), 3 items are selected with the provision that carrier phrase lengths of 4–6 syllables and embedding positions (start, mid, high) are balanced across all target word lengths and lexical frequency classes (Lehner & Ziegler, 2021). Items are presented in a randomised order.

Target word		carrier phrase			
L	F	L	P	Example 1	Example 2
1	High	4	Medial	Hier steht nur 'Arzt' drauf. *It just says ‚doctor' on here.*	Er versteht 'Schlaf' falsch. *He misunderstands ‚sleep'.*
1	Mid	6	Medial	Konnten Sie ' Pfad ' verstehen? *Could you understand 'path'?*	Jetzt ist 'sanft' an der Reihe. *Now it's the turn of 'soft'.*
1	Low	5	Final	Vorher kommt nur noch 'Gruft'. *Before that, there is only 'crypt'.*	Das Beispiel hier heißt 'Spind'. *The example here is called 'locker'.*
2	High	5	Initial	'Halsband' ist gar nicht so leicht! *'Collar' is not so easy*!	'Wundern' steht im Lexikon. *'Wondering' is in the dictionary.*
2	Mid	5	final	Verstehst Du das Wort 'Erdnuss'? *Do you understand the word 'peanut'?*	Mit Nummer acht folgt 'edel'. *It follows 'noble' with number eight.*
2	Low	4	Medial	Und hier steht 'Luftweg' drauf. *And this says 'airway' on it.*	Wir klicken 'Leinöl' an. *We click on 'Linseed oil'.*
3	high	6	Initial	'Weihnachtsbaum' steht hier als siebtes Wort. *'Christmas tree' is the seventh word here.*	'Meeresgrund' ist doch nicht so einfach! *'Seabed' is not so simple after all!*
3	Mid	5	Medial	War denn 'Trockenfleisch' schon dabei? *Was 'dried meat' already on the list?*	Da steht 'Weinprobe' gut lesbar. *It says 'wine tasting' quite legibly.*
3	Low	4	Final	Was reimt sich auf 'Langläufer'? *What rhymes with ' cross-country skier'?*	Das Wort heißt doch 'Wagendach'! *But the word is 'wagon roof'!*

L: length [syll]; **P**: embedding position; **F**: lexical frequency.

illustration of two parallel lists (excerpts) that could have been chosen for a patient's test and retest examinations). Hence, study 2 tests whether the stimulus selection process implemented in *KommPaS* yields stable intelligibility estimates.

The speech samples of the second examinations were placed on the microtasking platform with a time delay of several weeks from the first run, with the aim of simulating a pre-post period situation typical for treatment monitoring, and to decrease the likelihood that overlapping listener panels were recruited for corresponding test-retest pairs. Considering that the two tests were evaluated by different crowdworker panels, study 2 also tests whether the listener selection process implemented in KommPaS yields stable outcomes.

Statistical analysis

Retest reliability coefficients were calculated for all variables using intraclass correlation models (*two-way random effects, absolute agreement, single measurement*; cf. Koo & Li, 2016). In a second analysis we used paired t-tests to examine whether there were any systematic sequence effects.

Results

In Figures 3 and 4, the *KommPaS* scores of the two runs are plotted against each other. Figure 3 depicts each of the four KommPaS parameters in a separate panel. The figures illustrate that in all cases the scores were close to the diagonal (solid line), indicating a high agreement.

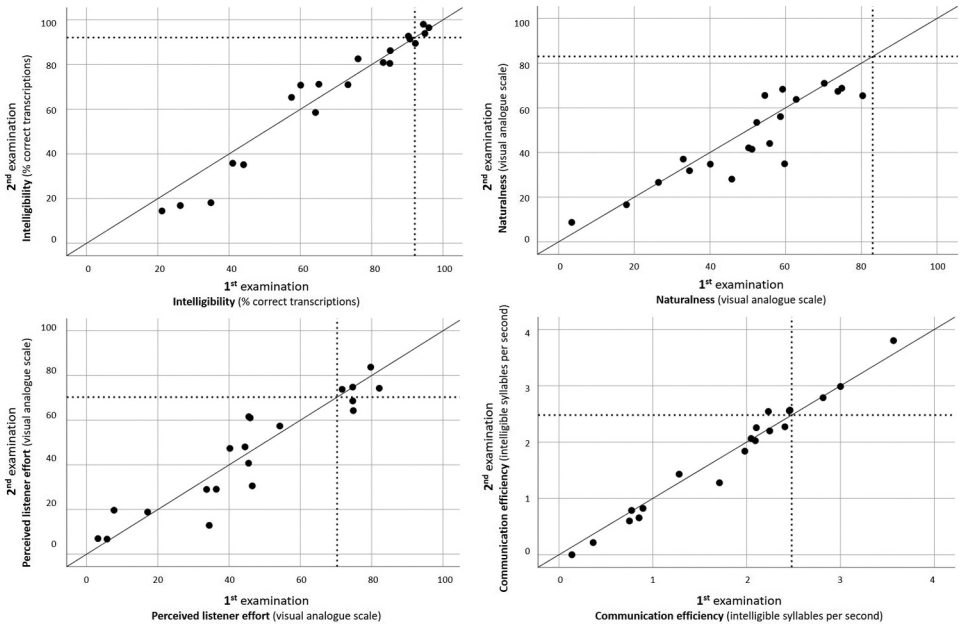

Figure 3. Test- vs. retest results for the four *KommPaS* parameters in 20 patients with dysarthria of varying degrees of severity. The dashed lines represent the 5% cutscores determined in study 1.

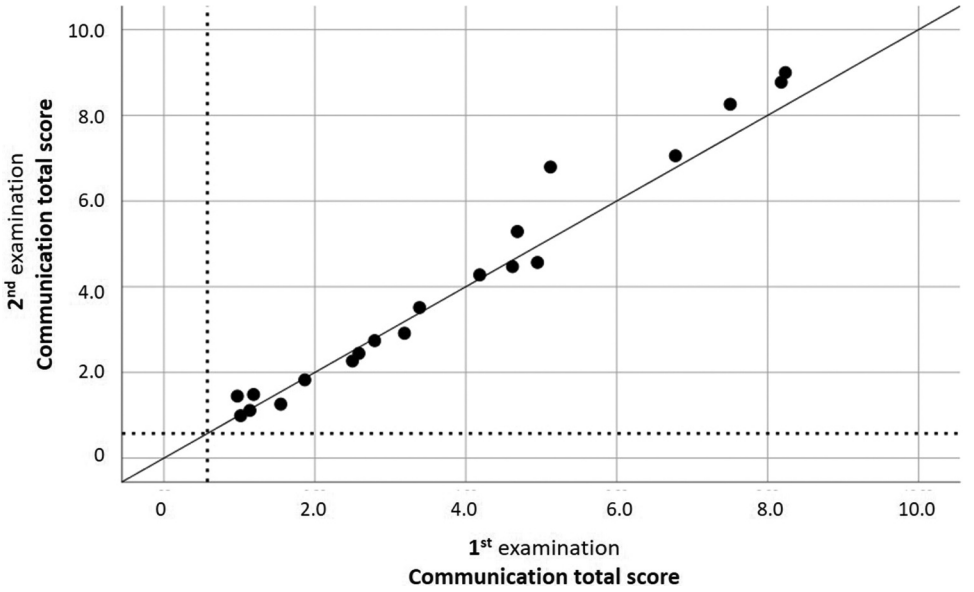

Figure 4. Test-retest agreement of the *communication total score* (as in Figure 3). The dashed lines represent the cutscore determined by a ROC analysis in study 1.

Table 5. Retest reliability of the *KommPaS* variables in terms of intraclass correlation coefficients (ICC) and pairwise comparisons of means (n = 20).

	ICC (2, 1) [95% CI]*	Difference (test – retest) [*p*-value; sig.]
Intelligibility	0.97 [0.92– 0.99] (excellent)	1 [0.38; n.s.]
Naturalness	0.88 [0.71– 0.95] (good)	4 [0.06; n.s.]
Perceived listener effort	0.93 [0.83– 0.98] (excellent)	0 [0.84; n.s.]
Communication efficiency	0.98 [0.96– 0.99] (excellent)	0 [0.54; n.s.]
Communication total score	0.98 [0.95– 0.99] (excellent)	– 0.2 [0.08; n.s.]

*Labels according to Koo and Li (2016)

The same was true for the total score (Figure 4). Note that the orientation of the *communication total score* is inverse to the four individual scales, with high scores indicating poor performance.

Considering whether the allocation of a patient to either the 'unimpaired' or the 'impaired' range of the distributions differed between the first and the second test, Figures 3 and 4 reveal that this did not happen to any clinically relevant extent. Only in a few cases with scores close to the cutoff-values (dashed lines) a change of the classification from one side to the other occurred, though with extremely small differences between the first and the second assessment.

The corresponding intraclass correlation coefficients (ICC) summarized in Table 5 indicate good to excellent reliability for all variables. The rightmost column of Table 5 lists the differences between the mean scores obtained in the second vs. first testing and the *p*-values of the corresponding paired t-tests, demonstrating that the differences were small and there was no significant sequence effect in any of the five variables.

Discussion

Despite the imponderables of ad-hoc compilation of ever-new test materials and listener panels in the *KommPaS* test protocol, the five scales achieved good to excellent retest reliabilities. In addition, there was no significant sequence effect between the two successive tests for any of the five parameters. Due to the overall small differences between first and second testing, changes in the classification of test scores as positive or negative, respectively, occurred only in very few cases, i.e., when a score was close to the cutoff and a small difference sufficed to shift it across the threshold. Note that the continuous metrics used in *KommPaS* can deal with such borderline cases in a clinically meaningful way, whereas a pure classification account would diagnose them as impaired or unimpaired, depending on only marginal differences.

Through its high retest reliability, the *KommPaS* test format ensures that noticeable changes in the test profile reflect meaningful changes in performance, e.g., due to disease progress, therapy induced improvements, application of a compensation strategy, or fatigue.

General discussion and clinical implications

The two studies presented here dealt with a novel approach to measure communication-related parameters of dysarthria by a tool that is suitable for clinical standard diagnostics. The new method is implemented as a web application, named *KommPaS*, that is accessible online for SLTs in clinics and private practices.

The test-to-test variation of listener panels and test materials that is central to the *KommPaS* approach raises important psychometric issues, i.e., if the method is accurate enough to distinguish between speakers with dysarthria and neurologically healthy speakers (study 1), and if it is stable enough to yield similar scores if the same patient is assessed two times in immediate succession (study 2).

In study 1, cutscores determined to keep the rates of false positive classifications below 5% left considerable proportions of patients with dysarthria classified as unimpaired regarding *intelligibility, perceived listener effort* and *communication efficiency*. Recall that for these variables only a moderate sensitivity was predicted for obvious clinical reasons. In contrast, *naturalness* showed a high sensitivity to dysarthric impairment and can therefore be taken as a proxy for dysarthria as viewed from the layperson's perspective, with a classification accuracy of 0.94. The crowdworkers' conception of unnatural speech was apparently broad enough to encompass a wide variety of perceptual dysarthria features accessible to their ears, but at the same time sufficiently insensitive to the variability inherent in typical speech to ensure a high discriminatory accuracy. The *communication total score* achieved a still higher known-groups validity of 0.96 for dysarthria vs. neurotypical speech, with an almost even balance between specificity and sensitivity. Hence, despite the imponderables of the crowd-based listener management and the ad-hoc compilation of ever-new speech materials, two of the five *KommPaS* variables proved very satisfactory in distinguishing between speakers with and without dysarthria. More importantly, central tendency and dispersion measures determined for the neurotypical participants of study 1 gave rise to standard metrics describing a speaker's distance from normal performance on each of the five parameters. Such metrics are useful for the assessment of the course of communication limitations in disease progression, recovery, or under treatment.

Study 2 revealed that the approach also has a high retest reliability. Obviously, the systematic control of lexical and length parameters in the random selection of test words and carrier phrases (Lehner & Ziegler, 2021) and the proposed method of aggregating the crowdworker scores ensured a high stability of test scores in repeated *KommPaS* tests of the same patient. The high retest reliability of the assessment proposed here guarantees sufficiently small measurement errors.

Conclusion, limitations and future work

The present work relates to a novel online diagnostic approach to assess communication-related features of dysarthria for clinical and research purposes. Motivated by theoretical considerations regarding the factors that influence the perception of such features, the proposed tool is based on principles of 'controlled variation' in the recruitment of listeners and the compilation of test materials. The strategies to avoid the psychometrically detrimental effects of the methodologically driven variation proved effective in that they ensured (1) a good overall separation of typical from dysarthric speech and (2) a high test-retest reliability.

The test is flexible enough to keep a close check on the natural course of dysarthric impairment or the effectiveness of treatment in daily clinical practice, and thanks to the high efficiency of crowdsourcing, it can process a large number of tests within short processing times.

Although this method of listener recruitment is very promising, there are some important points to consider with regard to its clinical application. First, special precautions must be taken for the processing of speech samples by crowdworkers in order to protect patient data as described in Lehner et al. (under review). Secondly, since the evaluation of speech samples by crowdworkers is associated with costs (see above), the use of KommPaS depends on the willingness of the clinical institution to pay for this external diagnostic service, just as they usually do for clinical chemistry or other laboratory services

The *KommPaS* parameters provide therapeutically useful information with a high face validity for how patients with dysarthria are perceived from a non-expert 'real-world perspective'. However, as they measure such a complex trait as verbal communication through highly constrained elicitation and evaluation methods, further validation studies are needed to establish the content and construct validity of this diagnostic tool. Investigations of the relationship of the *KommPaS* parameters with patient-reported outcome measures and with expert-based measures of dysarthric impairment have been completed and await being reported.

Future work is also required to expand the sample sizes. *KommPaS* will be developed as a test based on standard norms derived from a larger and more representative sample of patients with dysarthrias of diverse etiologies. In the course of this extension, the group of healthy control participants will also be expanded to establish more robust cutscores. A further important amendment is to demonstrate that the test is sensitive to change.

Acknowledgments

This work was funded by the German Academic Scholarship Foundation and the Bayerische Sparkassenstiftung. We are grateful to Jakob Pfab and Klaus Jänsch for web development and technical support. We wish to thank all patients, volunteers, and anonymous crowdworkers for participating in this research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by a PhD fellowship from the Studienstiftung des Deutschen Volkes awarded to the first author and a grant from the Bayerische Sparkassenstiftung.;

Ethics

Ethical approval (Project No. 19-365) was obtained from the Ethics Committee of the Faculty of Medicine at the Ludwig-Maximilians-University Munich, Germany. Participants were informed in detail about the study by the examiner and gave their written consent to participate.

KommPaS Study Group (clinical collaborators)

Madleen Klonowski[b], Nadine Geißler[b], Franziska Ammer[c], Christina Kurfeß[c], Holger Grötzbach[c], Alexander Mandl[d], Felicitas Knorr[d], Katrin Strecker[d], Theresa Schölderle[d], Sina Matern[e], Christiane Weck[e], Berthold Gröne[f], Stefanie Brühl[g], Christiane Kirchner[g], Ingo Kleiter[h], Ursula Sühn[h], Joachim von Eichmann[i], Christina Möhrle[j], Pete Guy Spencer[j], Rüdiger Ilg[k], Doris Klintwort[k], Daniel Lubecki[l], Steffy Marinho[m], Katharina Hogrefe[m])

[b]Schön Klinik München Schwabing
[c]Asklepios Klinik Schaufling
[d]Stiftung ICP München
[e]Krankenhaus Agatharied
[f]Kliniken Schmieder Allensbach
[g]St. Mauritius Therapieklinik Meerbusch
[h]Marianne-Strauß-Klinik Berg
[i]m&i-Fachklinik Enzensberg
[j]Hegau-Jugendwerk Gailingen
[k]Asklepios Stadtklinik Bad Tölz
[l]NeuroKom Bad Tölz
[m]Mutabor e.V.

ORCID

Katharina Lehner http://orcid.org/0000-0002-5071-8112
Wolfram Ziegler http://orcid.org/0000-0002-5760-1232

References

Barreto, S. S., & Ortiz, K. Z. (2020). Speech intelligibility in dysarthrias: Influence of utterance length. *Folia Phoniatrica Et Logopaedica*, 72, 202-210. https://doi.org/10.1159/000497178

Beukelman, D., Childes, J., Carrell, T., Funk, T., Ball, L. J., & Pattee, G. L. (2011). Perceived attention allocation of listeners who transcribe the speech of speakers with amyotrophic lateral sclerosis. *Speech Communication*, *53*(6), 801–806. https://doi.org/10.1016/j.specom.2010.12.005

Beukelman, D., Gillespie, L., Fager, S., & Ullman, C. (2014). Perceived attention allocation of listeners who transcribe the speech of dysarthric speakers with traumatic brain injury. *Journal of Medical Speech-Language Pathology*, *21*(3), 261–266.

Beverly, D., Cannito, M. P., Chorna, L., Wolf, T., Suiter, D. M., & Bene, E. R. (2010). Influence of stimulus sentence characteristics on speech intelligibility scores in hypokinetic dysarthria. *Journal of Medical Speech - Language Pathology*, *18*(4), 9+.

Borrie, S. A., Lansford, K. L., & Barrett, T. S. (2017). Generalized adaptation to dysarthric speech. *Journal of Speech, Language & Hearing Research*, *60*(11), 3110–3117. https://doi.org/10.1044/2017_JSLHR-S-17-0127

Chiu, Y. F., & Forrest, K. (2018). The impact of lexical characteristics and noise on intelligibility of parkinsonian speech. *Journal of Speech, Language, and Hearing Research*, *61*(4), 837–846. https://doi.org/10.1044/2017_jslhr-s-17-0205

Clickworker. (2021). Retrieved February 21, 2021, from https://www.clickworker.com/clickworker-crowd/

Crocker, L. M., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Dagenais, P. A., Brown, G. R., & Moore, R. E. (2006). Speech rate effects upon intelligibility and acceptability of dysarthric speech. *Clinical Linguistics & Phonetics*, *20*(2–3), 141–148. https://doi.org/10.1080/02699200400026843

Dagenais, P. A., Watts, C. R., Turnage, L. M., & Kennedy, S. (1999). Intelligibility and acceptability of moderately dysarthric speech by three types of listeners. *Journal of Medical Speech-Language Pathology*, *7*(2), 91–95.

DePaul, R., & Kent, R. D. (2000). A longitudinal case study of ALS: Effects of listener familiarity and proficiency on intelligibility judgments. *American Journal of Speech-Language Pathology*, *9*(3), 230–240. https://doi.org/10.1044/1058-0360.0903.230

Duffy, J. R. (2020). Examination of motor speech disorders. In J. R. Duffy (Ed.), *Motor speech disorders. Subtrates, differential diagnosis, and management* (4th ed., pp. 57–89). Elsevier.

Dykstra, A. D., Hakel, M. E., & Adams, S. G. (2007). Application of the ICF in reduced speech intelligibility in dysarthria. *Seminars in Speech and Language*, *28*(4), 301–311. https://doi.org/10.1055/s-2007-986527

Enderby, P. (2011). The Frenchay Dysarthria Assessment. International Journal of Language & Communication Disorders, 15, 165-173. https://doi.org/10.3109/13682828009112541

Feenaughty, L., Tjaden, K., & Sussman, J. (2014). Relationship between acoustic measures and judgments of intelligibility in Parkinson's disease: A within-speaker approach. *Clinical Linguistics & Phonetics*, *28*(11), 857–878. https://doi.org/10.3109/02699206.2014.921839

Gurevich, N., & Scamihorn, S. L. (2017). Speech-language pathologists' use of intelligibility measures in adults with dysarthria. *American Journal of Speech-Language Pathology*, *26*(3), 873–892. https://doi.org/10.1044/2017_AJSLP-16-0112

Hustad, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language & Hearing Research*, *51*(3), 562–573. https://doi.org/10.1044/1092-4388(2008/040)

Janbakhshi, P., Kodrasi, I., & Bourlard, H. (2019). Spectral subspace analysis for automatic assessment of pathological speech intelligibility. In *Proc. Interspeech 2019* (pp. 3038–3042). https://doi.org/10.21437/Interspeech.2019-2791

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). Thieme.

Kent, R. D., & Kim, Y. (2010). The assessment of intelligibility in motor speech disorders. In A. Lowit & R. D. Kent (Eds.), *Assessment of motor speech disorders* (pp. 21–38). Plural Publishing.

Kent, R. D., Kent, J. F., Weismer, G., Martin, R. E., Sufit, R. L., Brooks, B. R., & Rosenbek, J. C. (1989). Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects. *Clinical Linguistics & Phonetics*, *3*(4), 347–358. https://doi.org/10.3109/02699208908985295

Klopfenstein, M., Bernard, K., & Heyman, C. (2020). The study of speech naturalness in communication disorders: A systematic review of the literature. *Clinical Linguistics & Phonetics*, *34*(4), 327–338. https://doi.org/10.1080/02699206.2019.1652692

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Laganaro, M., Fougeron, C., Pernon, M., Levêque, N., Borel, S., Fournet, M., Catalano Chiuvé, S., Lopez, U., Trouville, R., Ménard, L., Burkhard, P. R., Assal, F., & Delvaux, V. (2021). Sensitivity and specificity of an acoustic- and perceptual-based tool for assessing motor speech disorders in French: The MonPaGe-screening protocol. *Clinical Linguistics & Phonetics*, 1–16. https://doi.org/10.1080/02699206.2020.1865460

Lansford, K. L., & Liss, J. M. (2014a). Vowel acoustics in dysarthria: Mapping to perception. *Journal of Speech, Language, and Hearing Research*, *57*(1), 68–80. https://doi.org/10.1044/1092-4388(2013/12-0263)

Lansford, K. L., & Liss, J. M. (2014b). Vowel acoustics in dysarthria: Speech disorder diagnosis and classification. *Journal of Speech, Language, and Hearing Research*, *57*(1), 57–67. https://doi.org/10.1044/1092-4388(2013/12-0262)

Lehner, K., Pfab, J., & Ziegler, W. (in press). Web-based assessment of communication-related parameters in dysarthria: Development and implementation of the KommPaS web app. *Clinical Linguistics & Phonetics*

Lehner, K., & Ziegler, W. (2021). The impact of lexical and articulatory factors in the automatic selection of test materials for a web-based assessment of intelligibility in dysarthria. *Journal of Speech, Language, and Hearing Research*, *64*(6S), 2196–2212. https://doi.org/10.1044/2020_JSLHR-20-00267

MacDonald, M. C., & Hsiao, Y. (2018). Sentence comprehension. In S.-A. Rueschemeyer & M. G. Gaske (Eds.), *The Oxford handbook of psycholinguistics* (pp. 171–197). Oxford University Press.

McAllister Byun, T., Halpin, P. F., & Szeredi, D. (2015). Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders*, *53*, 70–83. https://doi.org/10.1016/j.jcomdis.2014.11.003

McAuliffe, M. J., Fletcher, A. R., Kerr, S. E., O'Beirne, G. A., & Anderson, T. (2017). Effect of dysarthria type, speaking condition, and Listener Age On Speech Intelligibility. *American Journal of Speech-Language Pathology*, *26*(1), 113–123. https://doi.org/10.1044/2016_AJSLP-15-0182

McAuliffe, M. J., Gibson, E. M. R., Kerr, S. E., Anderson, T., & LaShell, P. J. (2013). Vocabulary influences older and younger listeners' processing of dysarthric speech. *The Journal of the Acoustical Society of America*, *134*(2), 1358–1368. https://doi.org/10.1121/1.4812764

Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language & Communication Disorders*, *48*(6), 601–612. https://doi.org/10.1111/1460-6984.12061

Nightingale, C., Swartz, M., Ramig, L. O., & McAllister, T. (2020). Using crowdsourced listeners' ratings to measure speech changes in hypokinetic dysarthria: A proof-of-concept study. *American Journal of Speech-Language Pathology*, *29*(2), 873–882. https://doi.org/10.1044/2019_AJSLP-19-00162

RCoreTeam. (2020). *R: A Language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Scherer, K. R. (1972). Judging personality from voice: A cross-cultural approach to an old issue in interpersonal perception. *Journal of Personality*, *40*(2), 191–210. https://doi.org/10.1111/j.1467-6494.1972.tb00998.x

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 & Hearing Research*, *59*(2), 216–229. https://doi.org/10.1044/2015_JSLHR-S-15-0086

Schölderle, T., Staiger, A., Schumacher, B., & Ziegler, W. (2019). The impact of dysarthria on laypersons' attitudes towards adults with cerebral palsy. *Folia Phoniatrica Et Logopaedica*, *71*(5–6), 309–320. https://doi.org/10.1159/000493916

Selouani, S. A., Dahmani, H., Amami, R., & Hamam, H. (2012). Using speech rhythm knowledge to improve dysarthric speech recognition. *International Journal of Speech Technology*, *15*(1), 57–64. https://doi.org/10.1007/s10772-011-9104-6

Smith, C. H., Patel, S., Woolley, R. L., Brady, M. C., Rick, C. E., Halfpenny, R., Rontiris, A., Knox-Smith, L., Dowling, F., Clarke, C. E., Au, P., Ives, N., Wheatley, K., & Sackley, C. M. (2019). Rating the intelligibility of dysarthic speech amongst people with Parkinson's Disease: A comparison of trained and untrained listeners. *Clinical Linguistics & Phonetics*, 33 (10–11), 1063–1070. https://doi.org/10.1080/02699206.2019

Stipancic, K. L., Tjaden, K., & Wilding, G. (2016). Comparison of intelligibility measures for adults with Parkinson's disease, adults with multiple sclerosis, and healthy controls. *Journal of Speech, Language, and Hearing Research: JSLHR*, *59*(2), 230–238. https://doi.org/10.1044/2015_JSLHR-S-15-0271

Tjaden, K., & Wilding, G. (2011). Effects of speaking task on intelligibility in Parkinson's disease. *Clinical Linguistics & Phonetics*, *25*(2), 155–168. https://doi.org/10.3109/02699206.2010.520185

Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., & Ramig, L. O. (2012). Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. *IEEE Transactions on Biomedical Engineering*, *59*(5), 1264–1271. https://doi.org/10.1109/TBME.2012.2183367

Utianski, R. L., Lansford, K. L., Liss, J. M., & Azuma, T. (2011). The effects of topic knowledge on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria. *Journal of Medical Speech-Language Pathology*, *19*(4), 25–36. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3738182/

Vojtech, J. M., Noordzij, J. P., Cler, G. J., & Stepp, C. E. (2019). The effects of modulating fundamental frequency and speech rate on the intelligibility, communication efficiency, and perceived naturalness of synthetic speech. *American Journal of Speech-Language Pathology*, *28*(2S), 875–886. https://doi.org/10.1044/2019_AJSLP-MSC18-18-0052

Walshe, M., & Miller, N. (2011). Living with acquired dysarthria: The speaker's perspective. *Disability and Rehabilitation*, *33*(3), 195–203. https://doi.org/10.3109/09638288.2010.511685

Walshe, M., Peach, R. K., & Miller, N. (2009). Dysarthria impact profile: Development of a scale to measure psychosocial effects. *International Journal of Language & Communication Disorders*, *44* (5), 693–715. https://doi.org/10.1080/13682820802317536

Whitehill, T., & Wong, C. (2006). Contributing factors to listener effort for dysarthric speech. *Journal of Medical Speech-Language Pathology*, *14*, 335–342.

Yorkston, K., Beukelman, D., & Tice, R. (1996). *Sentence Intelligibility Test (SIT). Communication Disorders Software*. In Tice Technology Services, Inc.

Yorkston, K. M. (2010). *Management of Motor Speech Disorders in Children and Adults* (3rd ed.). Pro-Ed. https://books.google.de/books?id=SuNIPwAACAAJ

Yorkston, K. M., & Beukelman, D. R. (1981). Communication Efficiency of Dysarthric Speakers as Measured by Sentence Intelligibility and Speaking Rate. *Journal of Speech and Hearing Disorders*, *46*(3), 296–301. https://doi.org/10.1044/jshd.4603.296

Ziegler, W., & Lehner, K., & KommPaS Study Group. (2021). Crowdsourcing as a tool in the clinical assessment of intelligibility in dysarthria: How to deal with excessive variation. *Journal of Communication Disorders*, *93*, 106135. https://doi.org/10.1016/j.jcomdis.2021.106135

Ziegler, W., Schölderle, T., Staiger, A., & Vogel, M. (2018). *Bogenhausener Dysarthrieskalen (BoDyS) [Bogenhausen Dysarthria Scales (BoDyS)]*. Hogrefe.

Ziegler, W., Staiger, A., Schölderle, T., & Vogel, M. (2017). Gauging the Auditory Dimensions of Dysarthric Impairment: Reliability and Construct Validity of the Bogenhausen Dysarthria Scales (BoDyS). *Journal of Speech, Language & Hearing Research*, *60*(6), 1516–1534. https://doi.org/10.1044/2017_JSLHR-S-16-0336