**Clinical assessment of communication-related speech parameters in dysarthria:  
The impact of perceptual adaptation.**

Vera Wolfrum1, Katharina Lehner2; Stefan Heim3,4,5, Wolfram Ziegler2

1 Department of Neurology, Medical Faculty, RWTH Aachen University, Aachen, Germany

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

3 Department of Psychiatry, Psychotherapy, and Psychosomatics, Medical Faculty, RWTH Aachen,  
 Aachen, Germany;

4 Research Center Jülich, Institute of Neuroscience and Medicine (INM-1), Jülich, Germany;

5JARA – Translational Brain Medicine, Aachen, Germany;

Correspondence:

Wolfram Ziegler, EKN, Institute of Phonetics and Speech Processing, Schellingstraße 3,  
80799 München, Germany

E-mail: wolfram.ziegler@ekn-muenchen.de

The authors declare no conflict of interest.

Keywords: dysarthria; auditory-perceptual; assessment; communication; SLT; crowdsourcing

**Abstract**

*Purpose*: In current clinical practice, intelligibility of dysarthric speech is commonly assessed by speech-language therapists (SLTs), in most cases by the therapist caring for the patient being diagnosed. Since SLTs are familiar with dysarthria in general and with the speech of the individual patient to be assessed in particular, they have an adaptation advantage in understanding the patient’s utterances. We examined whether and how listeners' assessments of communication-related speech parameters vary as a function of their familiarity with dysarthria in general and with the diagnosed patients in particular.

*Method*: Intelligibility, speech naturalness and perceived listener effort were assessed in 20 persons with dysarthria (PWD). Patients’ speech samples were judged by the individual treating therapists, five dysarthria experts who were unfamiliar with the patients, and crowd-sourced naïve listeners. Adaptation effects were analyzed using (i) linear mixed models of overall scoring levels, (ii) regression models of severity dependence, (iii) network analyses of between-listener and between-parameter relationships and (iv) measures of intra- and inter-observer consistency.

*Results*: Significant advantages of dysarthria experts over laypeople were found in all parameters. An overall advantage of the treating therapists over non-familiar experts was only seen in listening effort. Severity-dependent adaptation effects occurred in all parameters. The therapists’ responses were heterogeneous and inconsistent with those of the unfamiliar experts and the naïve listeners.

*Conclusions*: The way SLTs evaluate communication-relevant speech parameters of the PWD whom they care for is not only influenced by adaptation benefits, but also by therapeutic biases. This finding weakens the validity of assessments of communication-relevant speech parameters by the treating therapists themselves and encourages the development and use of alternative methods.

# Introduction

Dysarthria is a neurogenic motor speech disorder manifested by impairments of speech breathing, voice, articulation, and prosody. (Duffy, 2020). The gold standard in clinical dysarthria assessment is to analyze the speech of PWD for characteristics reflecting speech subsystem impairments, e.g., short breath groups, breathy voice, imprecise articulation, flattened prosody etc. (Darley et al., 1975). This diagnostic approach, which is supported by standard assessment tools such as the Mayo Clinic Rating System (Duffy, 2020), the Frenchay Dysarthria Assessment (FDA-2, Enderby & Palmer, 2012), the MonPaGe battery (Laganaro et al., 2021) or the BoDyS (Ziegler et al., 2018), aims to help clinicians disclose the profile of a patient’s speech impairment, identify dysarthria types, design tailored treatment approaches and monitor the progression or recovery of the patients’ specific respiratory, vocal, or articulatory dysfunctions.

As a consequence of these characteristics, dysarthric speech is often difficult to understand, it can be perceived as irritating or unnatural in conversational interactions, and it may evoke negative social, emotional or cognitive attributions in the patients’ interlocutors (Klopfenstein et al., 2020; Lehner, Ziegler, et al., 2022b; Schölderle et al., 2019; Yorkston et al., 1996). Thus, dysarthria may entail serious limitations in communicative participation and therefore calls for effective therapeutic interventions and counselling. Suitable diagnostic measures of communication-relevant aspects of dysarthric impairment are needed to help clinicians define ecologically valid outcome goals and monitor the achievement of these goals and the course of speech impairment.

***Communication-related speech parameters***

Clinical and experimental research have focused predominantly on three communication-related measures of dysarthric speech, i.e., *intelligibility*, *perceived listener effort*,and *naturalness* (e.g., (Dagenais et al., 2011; Lehner, Ziegler, et al., 2022b; Nagle & Eadie, 2018). Among them, intelligibility - defined as the accuracy with which listeners understand a speaker’s utterances – is the most extensively studied parameter (Kent & Kim, 2011; Yorkston et al., 1992). Its high relevance as an index of a person's ability to communicate successfully is obvious, which may explain why methods for measuring intelligibility have been incorporated into various clinical assessment tools (c.f., Enderby & Palmer, 2012; Laganaro et al., 2021) or established as separate intelligibility tests (Lehner, Pfab, et al., 2022; Martens et al., 2010; Yorkston et al., 1996; Ziegler & Zierdt, 2008).

Because intelligibility is not necessarily compromised in dysarthria, especially in mild disorders, other dimensions of communication impairment are clinically relevant as well. For example, conversing with PWD, even when they are completely understandable, can be effortful, - a feature that has been quantified using ratings of *perceived listener effort* (PLE), or *ease of listening* (EOL) (Fletcher et al., 2022; Landa et al., 2014; Nagle & Eadie, 2018; Whitehill & Wong, 2006). Various studies have found that perceived listener effort is inversely related with intelligibility, with particularly high effort in listening to PWD with low intelligibility vs. lower effort in intelligible patients (e.g., Beukelman et al., 2011; Landa et al., 2014; Whitehill & Wong, 2006)). The two parameters have a curvilinear relationship, as PWDs with intelligibility scores near 100% can vary greatly in the effort required to understand them, and PWDs scored at near-maximal listening effort can vary over a considerable range at the lower end of the intelligibility scale (Lehner, Ziegler, et al., 2022b).

*Naturalness* is conceived as a parameter that represents the degree to which a person’s way of speaking matches listeners’ expectations. It describes the “acceptability” of an individual’s speech characteristics, how “pleasant” it sounds to listeners, or, in negative terms, whether it causes discomfort or sounds “bizarre” (Dagenais et al., 2011; Darley et al., 1975; Klopfenstein et al., 2020; Strömbergsson et al., 2021). The impression of unnatural speech can be evoked by deviations in any speech dimension, i.e., respiration, voice, articulation, resonance, or prosody, which makes it a highly sensitive indicator of speech impairment in adults (Lehner, Ziegler, et al., 2022b; Schölderle et al., 2016; Ziegler et al., 2023).

Although naturalness and perceived listener effort are highly correlated with each other and with intelligibility, the three parameters are by far not redundant. For example, perceived listening effort is strongly influenced by intelligibility, but is more sensitive than measures of intelligibility in PWD with mild impairments. Naturalness, on the other hand, is a construct that is strongly influenced by voice and prosody parameters, while intelligibility and listening effort depend mainly on articulatory accuracy (for details see Lehner, Ziegler, et al., 2022b).

Despite their construct validity as communication parameters that add complementary information to intelligibility, perceived listener effort and speech naturalness have not been used in clinical standard assessment so far. One exception is the German *KommPaS* WebApp, which uses crowdsourcing as a tool to engage lay people as listeners in clinical diagnostics of these parameters (Lehner, Pfab, et al., 2022; Lehner & Ziegler, 2021; Lehner, Ziegler, et al., 2022a, 2022b; Ziegler et al., 2021).

***Familiarity and adaptation***

The auditory-perceptual assessment of features directly related to motor subsystem dysfunctions, such as respiratory, voice, or articulation problems, requires clinical experience and expertise as well as analytic listening skills, i.e., the ability to listen for phonetic form rather than meaning (Cucchiarini, 1993). These are prerequisites for extracting information about the type and severity of the disorder, recognizing key problems in terms of specific subsystem malfunctions, as well as deriving tailored treatment and individual counselling approaches (e.g., Pernon et al., 2022; Stipancic et al., 2021). In contrast, parameters such as intelligibility, naturalness, or perceived listener effort are of a different kind: They pertain to the domain of everyday communication and are accessible to everybody, - not only professional caregivers. No special phonetic or therapeutic training is needed to experience intelligibility loss, discomfort, or effort when listening to PWD. On the contrary, SLTs' therapeutic mindset and familiarity with the dysarthria characteristics of the patients they care for may even get in the way of recognizing how their patients' speech is perceived by others outside the therapy room.

There is convincing empirical evidence that the evaluation of communication-related parameters is strongly influenced by the degree to which listeners are familiar with or adapted to the speech characteristics of PWD (D'Innocenzo et al., 2006; Dagenais et al., 1999; DePaul & Kent, 2000; Tjaden & Liss, 1995). A large part of the experimental research done on this issue in the last decade addressed strategies to train listeners’ perception to better understand PWD and thereby reduce the everyday intelligibility burden of those affected (Borrie & Lansford, 2021; Lansford et al., 2023). Drawing on theoretical work on perceptual adaptation in speech processing (c.f., Kleinschmidt & Jaeger, 2015), two types of adaptation are distinguished in this research, i.e., adaptation to dysarthric speech in general, termed “group-specific adaptation” (Borrie et al., 2021), and adaptation to a particular individual’s pattern of dysarthria, termed “speaker-” or “talker-specific adaptation” (Borrie et al., 2012). To account for the particular clinical context of the present study, group-specific adaptation will in the following be referred to as “disorder-general” and speaker-specific adaptation as “patient-specific”.

*Disorder-general adaptation* exists in trained SLTs, especially those experienced in the assessment and treatment of dysarthria: They are familiar with the different speech patterns that can occur in dysarthria and have acquired, through training and experience, perceptual models of how specific symptoms such as reduced articulator movements, velar insufficiency or altered speech rhythm influence the acoustic signal. This gives them an advantage in parsing the acoustic signal and recognizing words and may thereby improve their understanding of dysarthric speech and reduce their effort in listening to PWD (e.g., Dagenais et al., 1999; Kleinschmidt & Jaeger, 2015; Lansford et al., 2023; Liss et al., 2002). According to a study by Borrie and coworkers, the disorder-general advantage of SLTs over laypeople increases with the amount of dysarthria-experience (Borrie et al., 2021).

*Patient-specific* adaptation occurs for example in a patient’s partner or relatives, who, in the course of constant communicative interactions, develop a perceptual model of the patient’s manner of speaking which increasingly supports their processing of the patient’s utterances (DePaul & Kent, 2000; Landa et al., 2014; Tjaden & Liss, 1995). This effect, which conforms to anecdotal evidence that even severely impaired PWD often communicate rather efficiently with their spouses, is compatible with more general mechanisms of perceptual learning and adjustment to unfavourable listening conditions or new dialects (Eisner & McQueen, 2005; Giroud et al., 2023; Kleinschmidt & Jaeger, 2015; Kraljic & Samuel, 2007; Lansford et al., 2023). Patient-specific adaptation in a patient’s relatives has been shown to increase with increasing familiarity over time (DePaul & Kent, 2000).

Patient-specific adaptation also occurs in clinical caregivers as an effect of their regular and intensive hands-on experience and therapeutic interaction with the patients they care for. Evidently, an SLT's adaptation to their client's speech, unlike that of a spouse or friend, is also driven by acquired analytic listening skills and knowledge of the mechanisms that create perceptual ambiguities in the speech signal of the PWD they care for (see below). Thus, whereas laypersons implicitly gain perceptual advantage through communicative interaction, patient-specific adaptation in SLTs is mediated, at least to some extent, through informed, analytical listening and explicit expertise. It complements existing disorder-general adaptation and may thereby reinforce the benefits in processing the speech of the patient they care for.

Most of the evidence collected on adaptation issues so far relates to intelligibility and supports the reasonable assumption that a higher degree of exposure leads to a greater advantage in understanding a patient's utterances. Regarding *perceived listener effort*, a similarly straightforward expectation is that familiarization with dysarthria in general or with the speech of a specific PWD continuously reduces listening effort. In fact, such effects have been demonstrated by Landa and coworkers, who found that teachers of children with dysarthria perceived significantly lower listening effort than unfamiliar listeners (Landa et al., 2014). As regards the perceived *naturalness* of dysarthric speech, little is known about potential adaptation effects. In a study of four speakers with and without dysarthria, Dagenais et al. (1999) found that SLTs rated the “acceptability” of unimpaired speakers significantly more harshly than naïve listeners and concluded that SLTs’ naturalness or acceptability construct might differ from that of laypeople. In contrast, in a study of alaryngeal speech by Finizia et al. (1998), SLTs rated acceptability more leniently than naïve listeners. Hence, the question remains to what extent naturalness ratings differ as a function of listeners’ professional experience with PWD.

***Clinical considerations***

From the patients' perspective, it is an immense advantage that those closest to them adapt to their speech, because it removes communication barriers between PWD and their partners, relatives and friends (DePaul & Kent, 2000). Listener-targeted perceptual learning strategies are being developed to enhance adaptation in close interlocutors and thereby reduce the burden of dysarthria in everyday communication (Borrie & Lansford, 2021).

However, from a diagnostic perspective, adaptation may have detrimental consequences for the reliability and validity of communication-related speech parameters assessed by SLTs, particularly by those responsible for the patient being diagnosed. The combined effects of disorder-general and patient-specific adaptation in SLTs could lead to gains in intelligibility and perceived listening ease relative to laypersons and create an expert concept of speech naturalness that is not representative of everyday communicative adequacy outside the clinic. Moreover, considering that the advantages of familiarity evolve over time (DePaul & Kent, 2000), follow-up assessments by therapists who care for a patient over longer periods of time are confounded by adaptive changes in their auditory-perceptual processes. Any therapy-related improvement in intelligibility or speech naturalness and any reduction in perceived listener effort may be due to either patient improvements or examiner adjustment, or both.

In the catalogues of dysarthric speech characteristics, intelligibility (and less frequently, naturalness/bizarreness) is listed on par with features reflecting dysfunctions of speech breathing, voice, articulation, and prosody (Darley et al., 1975; Enderby & Palmer, 2012; Laganaro et al., 2021). Hence, their assessment, insofar as it is a part of clinical diagnostics at all, is routinely performed by the same SLTs who also examine the patients’ speech motor problems, have regular contact with them, and are ultimately responsible for planning and implementing therapeutic interventions (Gurevich & Scamihorn, 2017). Standard clinical assessment of communication-related speech parameters is therefore not protected from confounding influences of disorder-general and patient-specific adaptation, mainly because the necessary infrastructural conditions for fast and inexpensive access to lay people as listeners are not given. Hirsch et al. (2022) recently investigated whether such confounds could in fact compromise the validity of conventional clinical intelligibility estimates. They found that SLTs' intelligibility ratings predicted lay people's intelligibility ratings relatively well, which, contrary to existing evidence for disorder-general adaptation effects, would support the credibility of common intelligibility assessment practice. However, methodological issues in this study warrant further investigation of the question of how SLTs’ perceptual adaptation impacts clinical evaluation of communicative speech parameters. For example, the study by Hirsch et al. (2022) was based on speech samples of a narrative text, namely the grandfather passage, which may have caused undesirable top-down effects on intelligibility (e.g., Hustad, 2007), especially as most SLTs are familiar with this particular text word for word. Moreover, the therapists’ intelligibility scores were based on subjective ratings, - a method that is known to be confounded with extraneous traits, such as unnatural speech or the experience of increased listening effort (Dagenais et al., 2006). Therapists who are asked to rate the intelligibility of a patient’s reading of the grandfather passage are likely to resort to extrinsic criteria, because they know what is being said anyway and will therefore need to speculate about how well others might understand the text. In addition, because Hirsch et al. (2022) examined therapists who were unfamiliar with the participating patients, the study did not address the most common clinical practice, i.e., intelligibility assessment by the therapist in charge.

***Study design, research questions and hypotheses***

The present study aimed to systematically assess disorder-general and patient-specific effects in the assessment of three communication related parameters, i.e., intelligibility (INT), perceived listener effort (PLE), and naturalness (NAT). To avoid validity issues, intelligibility was measured using transcription. In accordance with earlier studies, perceived listener effort and naturalness were assessed using visual-analogue scales (Klopfenstein et al., 2020).

We collected data from three types of listeners, i.e., (i) the SLTs who were caring for each individual patient, in the following referred to as *therapists*, (ii) SLTs experienced with dysarthria assessment and treatment who were unfamiliar with the enrolled PWD, in the following referred to as *experts*, and (iii) naïve listeners recruited through crowdsourcing (Lehner, Pfab, et al., 2022), henceforth referred to as *crowd*. [[1]](#footnote-1) Disorder-general adaptation effects were assessed for each variable by comparing naïve crowd listeners with experts, patient-specific adaptation by comparing therapists with experts. Differences between listener types were analyzed through linear mixed models. As familiarity effects were expected to depend on the severity of impairment, regression models including severity as predictor were calculated in a second step. Third, a network model was calculated to elucidate the correlation structure of the data. Finally, within- and between-listener consistency of response patterns were analyzed to examine structural similarities and dissimilarities between the responses of the three listener types.

Based on existing evidence of disorder-general and patient-specific effects on intelligibility, we hypothesized that the expert listeners would achieve overall higher scores than the crowd listeners, and the therapists higher scores than the experts. Given the evidence that listening effort is also modulated by adaptation, we expected a similar result for PLE, i.e., lowest effort in the therapists and highest in the crowd. Because ceiling effects predict that only little benefit will occur in patients with mild impairment, it was hypothesized that all effects are smaller in these patients. Expectations were less clear for the naturalness ratings. On the one hand one might expect that the adaptation benefits hypothesized for INT and PLE carry over to NAT, given the high covariation between these variables (Lehner, Ziegler, et al., 2022b). Moreover, frequent exposure of experts and therapists to dysarthric speech may make it sound more natural to them. On the other hand, NAT ratings could also be influenced by the professional awareness these listeners have of a patient’s complex speech profile, which might lead to an underestimation of naturalness by therapists and experts relative to the crowd listeners.

Regarding the multivariate correlation structure of the data and the consistency of response patterns, no predictions could be made in advance. The question here was whether we could identify one of the listener types as an “outlier” whose assessment criteria differed from those of the other two.

# Method

Ethical approval of this study has been obtained from the Ethics Committee of the Faculty of Medicine at the Ludwig- Maximilians-University Munich, Germany (Project No.19-365). All PWD and all listeners gave informed consent before taking part in the study.

***Speakers***

The study included 20 adult PWD (8 women, 12 men; 49.8 ± 4.4 years) recruited from cooperating clinics and private practices. All speakers met the following criteria according to the treating therapist: (a) a diagnosis of dysarthria, (b) age between 18 and 80 years (c) German as first language (d) no structural damage of the respiratory, laryngeal, or articulatory organs, (e) no associated aphasia or apraxia of speech; and (f) no manifest dementia. Table 1 shows the distribution of etiologies, age, and gender.

## ***- insert Table 1 about here -***

***Acquisition of speech samples***

The PWD were examined in a quiet room of the responsible clinic or private practice or their homes by the first author. In three cases, PWD were examined by an instructed colleague of the treating therapist. This ensured that the test items were unknown to the therapist who was caring for the PWD. A Sennheiser SC 40 USM monaural headset microphone or a Røde SmartLav Lavalier microphone and a laptop or tablet with a Wi-Fi connection were used as recording equipment.

Recordings were made following the KommPaS protocol (Lehner, Pfab, et al., 2022), which is described in short in the following.

Each PWD read aloud 30 sentences, which were presented one after the other on the computer screen. In patients with reading problems, the sentences were read aloud by the examiner and repeated by the PWD. Each examination started with the same three sentences followed by a set of 27 pseudo-randomly selected sentences, each of which consisted of a target word embedded in a syntactically and semantically neutral carrier sentence. For this purpose, an automatic selection was made from a set of more than 500 carrier phrases of four to six syllables in length, as well as target words from the *subtlex-np* corpus which comprises 12,829 one- to three-syllabic German content words. In the KommPaS WebApp the selection of stimulus sets is controlled for word frequency, sentence length, and initial, medial or final embedding position (Lehner & Ziegler, 2021). Participants were instructed to speak at their habitual rate and loudness while reading the sentences. A separate recording was made for each sentence. Trials were retaken in cases of reading or repetition errors.

After a manual check for recording quality, the speech samples, blocked per speaker, were placed on the German microtasking platform *Clickworker[[2]](#footnote-2)* utilizing an encrypted URL (Lehner, Pfab, et al., 2022).

***Listeners***

The study involved listeners of three different types. None of the participating listeners was informed about the goals and hypotheses of the study.

*Treating therapists (“therapists”)*

Consistent with common clinical practice, the SLTs who cared for the participating speakers served as listeners in evaluating the recordings of their respective patients. Ten SLTs (9 women) were involved, some of them responsible for more than one of the PWD and therefore acting as listeners several times. All treating therapists had a professional degree in speech and language therapy which qualified them for the treatment of dysarthria. At the time of their participation in this study, each therapist had been treating the corresponding PWD for at least fifteen therapy sessions, which was assumed as a realistic minimum for familiarization with the patients’ individual speech characteristics.

*Expert listeners (“experts”)*

A total of five listeners (4 women, 1 man; age 36 – 70 years) were recruited from an intern database for clinical experts. Each of them had clinical experience in assessment and treatment of PWD of at least ten years and a research background. The expert listeners evaluated the recordings of all participating PWD and received remuneration (€100) in exchange for their participation.

*Crowd-sourced naïve listeners (“crowd”)*

A total of 100 naïve listeners participated in 180 listening sessions. Following the KommPaS standard protocol, each of the 20 KommPaS sessions was completed by nine naïve listeners recruited through online crowdsourcing via the German microtasking platform *Clickworker* (see section *Scoring* below). Participation criteria as stipulated in the KommPaS standard protocol were (a) no hearing impairment, (b) German native speakers, and (c) no experience in communicating with people with neurological impairments (all criteria self-reported). As the KommPaS WebApp does not preclude listeners from taking part in more than one listening session, the number of listeners (n=100) was smaller than the number of listening sessions (20 [PWD] x 9 [listeners] = 180). Crowd listeners received € 1.75 per listening session. For privacy reasons, no demographic data are available for the crowd listeners.

***Listening sessions***

All listening sessions, including the therapists’ and the experts’ sessions, were administered following the *KommPaS* standard protocol via the microtasking platform *Clickworker* (Lehner, Pfab, et al., 2022). Listeners received individual links to the speakers’ recordings and completed the online listening tasks hosted on a dedicated KommPaS server. All listeners received the exact same written instructions.

Each listening session started with a brief introduction, reminding listeners to use headphones and to choose a quiet environment to complete the test. Listeners had to consent to the privacy policy, complete a short technical test of the headphones and a sample listening task. Thereafter, listeners were presented the recordings of the first three sentences of the KommPaS test materials to assess the perceived naturalness of the PWD’s speech on a visual analog scale (VAS) represented by a horizontal bar, with the anchors *very unnatural* [left, = 0] to *natural* [right, = 100]). To deflect their focus from the intelligibility of the spoken sentences, listeners were instructed that "the content of what is said is not important in this assessment” (Lehner, Pfab, et al., 2022).

For the assessment of intelligibility, listeners then listened to the 27 sentences of the intelligibility test, one by one. After each sentence, the corresponding written carrier sentence appeared on the screen after a pause of 1000 ms. The target word was replaced by a gap and was to be filled in by the listener using the computer keyboard. If unsure, listeners were asked to make their best guess or fill in a # sign if they felt unable to guess. Each sentence was presented only once, with no time limit for the response.

After completing the transcription task, listeners reported on another VAS the effort they had to expend to understand what was said (*high effort* [left, = 0] to *no effort* [right, = 100]). Note that the scaling of perceived listener effort assigned low numbers to high effort and vice versa, to align it with the NAT scale in the sense that severe impairment was allocated to the left, and mild or no impairment to the right end of the VAS.

Listening sessions lasted between 6 and 10 minutes. In an evaluation of ca 1,000 listening sessions, 95% of all sessions took less than 16 minutes (Lehner, Pfab, et al., 2022).

***Scoring***

For all listeners, the KommPaS WebApp automatically recognized the correctly transcribed target words, whereas spelling errors or homophones were marked manually as correct by the first author.

In a first step, a single score was determined for each listener type per speaker and parameter. For obvious reasons, the therapists’ scores relied on only one response per speaker and parameter, i.e., the ratings and transcripts provided by the patients’ individual therapists. Hence, as in standard clinical assessment, the therapists' data were collected on an *n=1* basis, with fixed speaker-listener assignments in each case. From the *expert* group, five datasets were obtained for each speaker and each parameter by the five expert listeners. Since there was no reason to believe that these listeners differed fundamentally in their commitment to the task, their scores were combined into arithmetic means per speaker and parameter. This method was justified post-hoc by an excellent internal consistency (see below). For the *crowd* listeners, scores were obtained according to the *KommPaS* protocol, i.e., from varying panels of nine naïve listeners each, assembled individually for each speaker from the *Crowdworker* platform on a first-come – first-serve basis. Because there was no experimental control of the composition of the crowd listener panels, their scores were combined using the weighted averaging method developed in Ziegler et al. (2021). This method assumes that among the nine listeners in a KommPaS panel, the one with the highest score in the transcription task must be the most reliable one in terms of attention and task commitment. This listener’s scores were weighted by 1, and all other listeners were given exponentially lower weights depending on the distance of their intelligibility scores from the “best” score. This method, which penalizes extreme downward outliers by low weights, has been shown to yield a high test-retest reliability and protect from “spammers”, i.e., inattentive listeners or listeners with a low task commitment, low attention, or poor technical equipment. A panel size of 9 was found optimal in terms of cost-benefit (Ziegler et al., 2021).

***Data analysis***

In a first analysis, differences between listener types were analyzed using linear mixed models, to determine the magnitudes of overall adaptation effects in the three parameters. Possible influences of severity on such effects were studied separately using second order polynomial regression models. The multivariate correlation structure of the 9 datasets (3 [parameters] x 3 [listener types]) was analyzed by exploratory graphical network modeling, with the aim of elucidating similarities and dissimilarities between the response patterns in the different listener types. In a final step, consistency analyses were performed to examine within- and between-listener consistencies. Statistical analyses and chart creation were performed using R (RStudio Team, 2022) and JASP (JASP Team, 2023). Further details on the specifications of the statistical analyses are provided at the appropriate places in the results section.

# Results

*Raw data*

Figure 1 plots the raw data for all listener types and all parameters, illustrating that the patients covered a broad range of severities. Regarding intelligibility, scores below 20% were rare, whereas in perceived listening ease and naturalness, the lower range of the two scales was more densely populated.

Figure 1 about here

As the data of the five expert listeners (E1 to E5) showed excellent inter-observer consistency in each parameter (**INT**: Cronbach’s *α* = 0.99, ICC(2,5) = 0.99; **PLE**: *α* = 0.97, ICC(2,5) = 0.95; **NAT**: *α* = 0.97, ICC(2,5) = 0.96), they were averaged across listeners to yield a mean expert score for each parameter. These scores will be used as a reference to determine disorder-general (experts vs. crowd) and patient-specific adaptation effects (experts vs. therapists). The averaged expert scores will also serve as an index of the severity of impairment in regression analyses of adaptation effects. In a final analysis step, the expert listener group will be disaggregated again to explore the consistencies of response sets based on single listeners.

*Overall adaptation effects*

Figure 2 presents marginal means and standard errors of the parameter scores for each listener type across the 20 PWD. On each parameter, the naïve listeners achieved overall lower scores than the dysarthria-adapted listeners (disorder-general adaptation), whereas a substantial patient-specific adaptation effect, i.e., therapists scoring higher than experts, was only seen in PLE.

Figure 2 about here

Type III ANOVAS using the Satterthwaite df-method were applied for fixed effects (R-package *lmerTest*, Kuznetsova et al. (2017); see Table 2 for model summaries). A significant disorder-general effect was obtained for INT, with ca. 8% lower intelligibility scores in the crowd vs. the expert listeners, on the average. The only significant patient-specific effect was seen in PLE. Combined group- and patient-specific effects were significant in all parameters, suggesting that at the group level the therapists overestimated the communicative value of the PWD’s speech relative to the naïve crowd listeners.

Likelihood ratio test statistics were used to test each model against a model without random effects (R-package *lmerTest*, function *ranova*; Kuznetsova et al. (2017)), yielding a significant influence of PWD as random-effects factor (Table 2). The nature of this influence will be determined in the next section.

Table 2 about here

*Severity-dependence of adaptation effects*

Considering that the margin for adaptation-induced gains is limited for speakers with mild impairments, ceiling effects may be expected. Moreover, because adapted and non-adapted listeners might tend to agree in patients with most severe dysarthria, floor effects could occur as well. Therefore, adaptation may not be uniformly distributed across severity scales. To detect potential severity influences on adaptation gains, difference scores between listener groups were modelled as a function of severity for each parameter, using the experts’ responses as severity references.

Figure 3 plots pairwise differences between listener types against the average scores of the expert listeners. Panels E-C (top row) and T-E (middle row) represent severity-dependent disorder-general- and patient-specific adaptation effects, respectively, while panels T-C (bottom row) depict combined effects. Positive values indicate effects going in the predicted direction. For example, in the E-C panels, differences > 0 indicate higher intelligibility scores in the experts as compared to the crowd listeners. Considering the curvilinear relationship seen in several panels, each data set was fitted in a first step by a 2nd order polynomial regression. Linear models were chosen only when the adjusted R2 value was greater in the linear than in the quadratic model. Table 3 presents summaries for all models.

*Intelligibility.* Regarding intelligibility (left column of Figure 3), linear regressions outperformed polynomial regressions in all comparisons. In all three diagrams, a linear trend was visible, indicating that the intelligibility gains of the adapted versus the non-adapted listeners were larger in PWD with lower intelligibility scores. In the experts vs. the crowd (panel E-C), the effect was small and non-significant (see Table 3), indicating a virtually uniform adaptation advantage of the expert listeners across the whole severity range. However, in the patient-specific (T-E) and the combined (T-C) adaptation models, significant severity effects were found. In the patients who had expert scores below 50%, the therapists’ intelligibility scores conformed to expectations, with gains of up to 20% relative to the experts and up to 30% relative to the crowd. In contrast, therapists scored surprisingly low on several patients in the higher intelligibility range, yielding unexpected negative difference scores in the T-E and the T-C panels. A look at the transcripts scored as errors in these cases showed that target words - in violation of instructions - were often transcribed by similar-sounding pseudowords, much like an attempt to reproduce the distorted sound pattern rather than a possibly intended word. This occurred even for items that were otherwise transcribed correctly by all crowd listeners and all experts. Examples of such errors are listed in Table A1 of the Appendix.

Figure 3 about here

*Perceived listener effort.* The model fits of the PLE difference scores in the E-C and T-C panels of Figure 3 represent 2nd order polynomial regressions, reflecting a tendency of greater adaptation effects in patients with PLE values in the middle range, and a convergence with the crowd scores towards the lower and the higher ends of the PLE scale. In contrast, the patient-specific adaptation advantage of the therapists compared to the expert listeners was highest for the severe cases and decreased linearly towards the upper end of the PLE range, similar to the pattern observed for intelligibility. In some of the PWD who caused considerable listening effort (experts’ PLE scores below 50), their corresponding therapists reported higher listening ease and exceeded the two other listener types by margins greater than 30. Table 3 reveals that only the E-C model became significant (R2=0.50, p < 0.01), whereas the two models depending on the therapists’ PLE scores missed the 5% significance level, obviously due to extremely high residual standard errors (Table 3, column RSE).

Table 3 about here

*Naturalness.* The shapes of the regression models plotted in Figure 3 (rightmost column) resembled those seen in PLE, with adaptation gains particularly in the middle range of the NAT scale and a negative linear trend in the patient-specific adaptation gains (i.e., T vs. E). Some of the PWD who sounded unnatural to the expert listeners, with NAT scores below ca. 50, were scored as considerably more natural by their individual therapists, with advantages of more than 25 scale points. In other cases, therapists’ NAT ratings were far lower than those of the experts or the crowd listeners. Table 3 reveals that only the E-C model reflecting the severity-dependence of the disorder-general adaptation of NAT ratings (i.e., experts scoring higher than crowd) became significant.

Note that the regression residuals of the models involving the therapists’ scores (i.e., T-E and T-C) had high standard errors in all three parameters (Table 3, column RSE), indicating a large variation of the therapists’ scores around their overall linear or curvilinear trends. The residual standard errors of the therapists’ deviations from the crowd and the expert scores exceeded those of the experts against the crowd by factors 2 to almost 4. The heterogeneity was largest in the two parameters involving subjective ratings, i.e., PLE and NAT, where it prevented significant model fitting.

**Correlation network of listener responses**

To answer the question of how listener responses were related between the three listener types and the three parameters, an exploratory psychometric network analysis of pairwise dependencies was conducted for the nine variables in the data set (Borsboom et al., 2021; Costantini et al., 2015). The model was based on an EPICglasso algorithm, which basically computes partial correlations between variables and applies a shrinkage algorithm to create a sparse matrix of connection weights by nullifying spurious correlations. In the graphical network model, variables are represented by nodes, and the edges connecting the nodes represent the weights of their LASSO-regularized partial correlations. The R package glasso with the tuning parameter set to 0.5 was used for model estimation (Friedman et al., 2019). Figure 4 presents a plot of the network, with connection weights < 0.2 suppressed (JASP Team, 2023). The full matrix of network weights is printed in the Appendix.

Figure 4 about here

Two features of this network are remarkable. First, nodes representing the three different parameters were relatively isolated from each other, suggesting that the INT-, PLE- and NAT-scales measured relatively independent traits. Second, the by far strongest connections in the network were between the crowd and the expert listeners within each of the three parameters, with weights of 0.69, 0.64, and 0.73, respectively, whereas the nodes representing the therapists’ responses were relatively isolated. Only the therapists’ intelligibility scores were weakly associated with the intelligibility scores of the crowd and expert listeners, though with a substantially lower weight of 0.34 in the network model

***Consistency of response patterns***

One might argue that the greater heterogeneity of the therapists’ responses compared to the crowd- and experts’ data (see Figure 3 and Table 3) and the weak association between patient-familiar and patient-unfamiliar listeners (Figure 4) was ascribable to the fact that the therapists’ responses were based on single and different listeners and therefore subject to between-listener variability. In contrast, the averaging of expert responses across five individual listeners and the KommPaS-weighting of crowd-sourced listeners’ responses (Ziegler et al., 2021) may have removed critical between-subject variation. To address this objection, consistency analyses were undertaken involving the five individual experts’ scores separately in combination with the crowd and therapists’ data.[[3]](#footnote-3)

In a first step, the regression models shown in Table 3 (models E-C and T-E) were re-designed so as to use the expert-based severity scores to predict the raw scores (rather than distance scores) of the therapists’ and the crowd listeners’ responses. These regression analyses were then expanded to also predict each single expert listener’s responses. To avoid part-whole correlation effects in these predictions, a Jackknife procedure was applied by excluding each predicted expert listener’s scores from the predictor. The residual standard error of these models was extracted as a measure of each listener’s intra-individual consistency adjusted for severity-dependent trends. All regression models were fitted by second-order polynomials to make the variability measures comparable across parameters, keeping in mind that linear models are a specific case of polynomials.

Figure 5 (left panel) plots these data for each parameter separately. Because the models were mathematically almost identical, the RSE-values listed in Table 3 (RSE-column, rows E-C vs. T-E) were replicated with high precision, yielding substantially lower residual standard errors in the crowd listeners than in the therapists’ data. Remarkably, the five expert listeners had low RSE values as well, resembling the crowd more closely than the therapists. In NAT and PLE, i.e., the two parameters measured by subjective ratings, the crowd listeners’ responses had the highest intra-individual consistency in terms of their variability of model residuals, while the therapists’ responses were clearly inconsistent with the rest.

Figure 5 about here

In a second step, the between-listener consistency of responses was examined by item-total-correlation analyses across the seven listeners (C, E1 - E5, T), i.e., computations of the correlations of each listener’s response pattern with the pattern averaged across the six other listeners. The right panel of Figure 5 plots these data for each of the three parameters. It demonstrates an overall high consistency of the results, with item-total-correlations > 0.9 in almost all cases. Notably, the crowd panels were among the most consistent responders, with item-total correlations comparable to the five expert listeners. In contrast, the therapists’ responses had the lowest consistency values, with particularly low item-total correlations for the NAT and the PLE ratings.

# Discussion

We will first briefly summarize the background of this study before discussing the data in terms of their evidence for disorder-general and patient-specific adaptation. Three further subsections then address issues of consistency and elaborate on the special position of treating therapists among the three types of listeners involved in assessments. We conclude by discussing the clinical consequences of our findings, with a focus on the common clinical practice that communication-related speech parameters are assessed by the patients’ individual caregivers.

***Background and hypotheses***

This study aimed to investigate how disorder-general and patient-specific adaptation affects the clinical assessment of communication-related parameters in PWD. To our knowledge, this is the first study comparing naïve listeners with dysarthria experts and with the patients’ individual therapists for their assessments of intelligibility, perceived listening ease, and speech naturalness. While the recruited naïve listeners were not familiar with dysarthria in general or with the patients studied here, the expert listeners were highly familiar with dysarthria through extensive therapeutic and research experience, but did not know the patients in this study. Finally, the therapist listeners were familiar not only with dysarthria through their professional training as SLTs, but also with the individual patients they assessed after at least 15 sessions of individual dysarthria therapy.

Evidence from studies of perceptual learning in speech processing (Eisner & McQueen, 2005; Kleinschmidt & Jaeger, 2015; Kraljic & Samuel, 2007) and, more specifically, processing of dysarthric speech (Borrie & Lansford, 2021; Borrie et al., 2021; D'Innocenzo et al., 2006; Dagenais et al., 1999; DePaul & Kent, 2000; Lansford et al., 2023; Tjaden & Liss, 1995) led us to predict that the expert listeners included here would have an intelligibility advantage over the lay persons (“disorder-general adaptation”). Furthermore, we hypothesized that the therapists, based on their extensive therapeutic experience with each diagnosed patient on top of their professional expertise, would have a further intelligibility advantage over the experts (“patient-specific adaptation”) and the lay persons (combined effect). Similar predictions were made for listening effort, considering that speaker familiarity is known to support speech processing in a rather general sense (Kleinschmidt & Jaeger, 2015). With respect to naturalness, expectations were less clear. Expertise and experience may, on the one hand, enhance listeners' sensitivity to all facets of dysarthric speech, which might compromise their sense of what laypersons perceive as natural and lead to underestimation of naturalness. On the other hand, to the extent that ratings of naturalness are influenced by intelligibility and listening effort (Lehner, Ziegler, et al., 2022b), the three parameters might be subject to concurrent adaptation effects (Strömbergsson et al., 2021).

***Disorder-general adaptation (experts vs. laypersons)***

Regarding intelligibility and perceived listener effort, the hypothesis of a disorder-general effect was strongly supported: The expert listeners’ transcription scores were ca. 8% higher than the laypersons’ scores, an effect that was significant and consistent across the whole intelligibility range (Figure 3, top left). Likewise, perceived listening effort was significantly lower (i.e., PLE scores were higher) in the experts, with an average advantage of more than 9 scale points. In PLE there was also a significant curvilinear dependence on severity, with virtually no difference between the crowd and the expert listeners at the low end of the scale, and an advantage of 15 scale points or more in terms of higher ratings around PLE-scores of ca. 70-75. At the upper end of the scale, the crowd listeners still perceived some effort where the experts already scored close to the maximum. The highest adaptation benefit was estimated here to occur in the third quartile of the PLE scale (Figure 3, top row, middle column).

A very similar relationship between the crowd and the expert listeners was observed for the naturalness ratings, though the linear mixed model of the whole data set missed significance. However, there was a systematic influence of severity on the amount of adaptation in the experts’ naturalness ratings, as shown again by a significant 2nd order polynomial curve. A maximum effect of more than 10 scale points occurred in the range between 50 and 75, whereas the advantage disappeared in the extremely severe and extremely mild cases (Figure 3, top row, right column).

In summary, the reported results confirm earlier findings that familiarity with dysarthric speech leads to a better understanding of PWD (e.g., Borrie et al., 2021; Borrie et al., 2012; Dagenais et al., 1999; Liss et al., 2002). Furthermore, we extended existing evidence to other communication-related parameters, i.e., perceived listening effort and naturalness, and quantified the dependence of familiarity effects on severity. Our results are also consistent with more general theories about implicit adaptation and perceptual learning in auditory speech processing under adverse conditions (Eisner & McQueen, 2005; Kleinschmidt & Jaeger, 2015; Kraljic & Samuel, 2007). This was not a matter of course, because clinical experts' adaptation to dysarthric speech is not necessarily the same as adaptation to a foreign dialect in everyday life. It is not exclusively based on implicit, passive familiarization, i.e., regular interactions in everyday communication, but is probably also mediated by professional training, recurring examination of dysarthric speech patterns in clinical practice, and insights into where the ambiguities in the speech signal come from. Hence, the explicitly acquired and phonetically specific perceptual skills of trained clinicians contribute to an implicit communicative advantage, i.e., better speech recognition with less effort.

***Patient-specific adaptation (therapists vs. experts)***

The analyses of patient-specific adaptation effects were dominated by large deviations of single therapists’ scores from the experts’ and the laypersons’ scores (Figure 3, leftmost column). Regarding intelligibility, deviations were observed in both directions: Conforming with expectations, several therapists had an intelligibility advantage of more than 10% over the experts and between 20% and 30% over the crowd listeners, especially in the lower intelligibility range. In contrast, others scored far lower than the experts and even lower than the crowd listeners, especially on patients in the intelligibility range between 70% and 100%. Therefore, there was no patient-specific adaptation in terms of an overall intelligibility advantage of the therapists over the experts and laypersons, but there was a clear and significant negative linear trend in the therapists’ adaptation gain, though with a large variation of regression residuals.

This trend was very similar in the PLE ratings, with the difference that overall higher values prevailed in the therapists, yielding a patient-specific adaptation effect in terms of a significantly lower overall listening effort, in addition to a significant influence of severity (Figure 3, middle column). In several patients in which experts and crowd listeners perceived high effort, their responsible therapists indicated much lower effort, with PLE gains (i.e., higher scores indicating lower effort) amounting to almost half of the scale width. At the upper end of the PLE scale, patient-specific adaptation gains were smaller or even absent, because of a ceiling effect. Overall, the finding of patient-specific PLE adaptation was consistent with a study by Landa et al. (2014), in which speech samples of children with cerebral palsy were rated as more easily understandable by school staff members who were familiar with the children than by unfamiliar listeners.

Regarding the naturalness ratings, no patient-specific adaptation was observed, neither in terms of overall mean values nor in severity-dependent effects (Figure 3, rightmost column). As in PLE, there was a similar negative linear trend in the adaptation gains of the therapists relative to the experts and the crowd listeners, but due to the large variation of regression residuals the regression model was not significant.

***Similarities and dissimilarities in listeners’ response patterns***

In an exploratory graphical network analysis of the multivariate dataset, the response patterns of the crowd- and the expert listeners were found to be structurally similar in terms of high covariation, despite the adaptation advantage of the latter (Figure 4). The INT-, PLE-and NAT-variables of the two listener types who were unfamiliar with the patients had by far the strongest pairwise connections in the network, whereas the therapists’ response variables were isolated. The close familiarity of the therapists with the dysarthria profiles of their individual patients obviously influenced their response behaviors in particular ways. The at least weak connection of their intelligibility scores with those of the experts and laypersons indicates that the transcription method of intelligibility assessment seemed to be more robust against such influences than the two parameters assessed through subjective ratings, i.e., PLE and NAT.

There was a clear separation between the three test parameters in the correlation network of Figure 4. This result is important from a psychometric perspective, as it confirms the discriminant validity of test profiles encompassing intelligibility, listening effort and naturalness as separate variables and underscores the value of assessing more communication parameters than just intelligibility. A similar finding has already been reported in a recent, more comprehensive study (Lehner, Ziegler, et al., 2022b).

***Variability and consistency***

A major result of the present study was that the therapists often assessed the communication-relevant speech parameters of their patients in idiosyncratic ways, deviating sometimes widely in both directions from the expert- and the crowd listeners, and also deviating widely from the linear or curvilinear trends that predicted adaptation gains along the severity continuum. While the regression models describing disorder-general advantages as a function of severity in experts vs. lay persons had narrow 95%-CI bands and small residual standard errors, the models of patient-specific adaptation (therapists vs. experts) had a much lower predictive value, with broad confidence intervals and large prediction errors.

The crowd- and the therapists’ data were comparable to the extent that each patient was assessed by a different listener (in the therapists) or a different listener panel (in the crowd). Nonetheless, the laypersons’ relationships with the experts’ mean scores as a reference were considerably less variable, which may have been due to the circumstance that, as a necessary test principle of KommPaS, the observation units in the crowd data are weighted averages of panels of nine listeners (Ziegler et al., 2021). To explore if the five single expert listeners, considered individually, were similarly inconsistent as the therapists, the mean expert scores were disaggregated and each single expert listener’s severity-adjusted response consistency was compared with that of the two other listener types. These analyses confirmed that the therapists’ inconsistency was in fact ascribable to their specific therapeutic familiarity with the patients rather than a general instability of single listener data. A second consistency analysis addressing between-listener agreements with disaggregated expert listener data confirmed the uniqueness of the therapists’ response patterns.

***The uniqueness of the treating therapists***

How can the special role of the therapists in the reported assessments be explained? At first glance, they had similar prerequisites as the experts, namely professional experience and trained analytic listening skills, plus a supposedly additive advantage resulting from their caretaker role, i.e., a prolonged period of therapeutic engagement with the patients, explicit reflections on the characteristics of their impairment and the therapeutic means by which it might be influenced, and probably also a monitoring of the changes they achieved in the course of the ongoing interventions. In several cases, especially at the lower end or in the middle of the three scales, their intensive therapeutic familiarity with the patients payed off in sometimes huge adaptation gains. In several other cases, however, only little or no patient-specific advantage of therapists over experts was seen, or even negative effects occurred. Some of the therapists who did not show the expected adaptation gain may have had difficulties distinguishing sharply enough between the judgment criteria in question, e.g., naturalness of speech or ease of listening, and their awareness of the respiratory, vocal, articulatory, and prosodic problems their patient was currently struggling with. Such confounds with extraneous concepts have been reported earlier, for example for subjective intelligibility ratings which can be blurred by speech features unrelated to the intelligibility criterion (Dagenais et al., 2006). A bias of this kind may explain why in some cases therapists judged considerably more harshly than the expert- and crowd listeners and, thus, against the expected adaptation benefit.

However, considering that the intelligibility measure used here was based on word transcription, these arguments can hardly explain why some of the therapists had lower intelligibility scores than the experts and even the crowd listeners, especially in patients with only mild or moderate intelligibility problems. Transcription is known to provide objective information about whether a listener understood what was said and to leave little room for subjective bias: As long as listeners comply with the instruction to write down what they understood, their transcript provides a valid measure of the accuracy of their decoding of the acoustic signal, regardless of the contributions of bottom-up or top-down processes that may have prompted their understanding (e.g., Dagenais et al., 2006; Lehner, Ziegler, et al., 2022b). This explains at least why the observed inconsistencies were much smaller in intelligibility than in naturalness and listening effort. Downward outliers in the accuracy of transcripts of dysarthric speech can occur as a consequence of fluctuating attention, poor quality of audio devices, or unfavorable listening conditions. Other potential sources of variability lie in listeners’ individual skills and perceptual strategies to exploit the ambiguous acoustic information in a degraded speech signal (Borrie et al., 2017; Kleinschmidt & Jaeger, 2015) or the extent of dysarthria-specific training and clinical experience with PWD (Borrie et al., 2021). However, as demonstrated through consistency analyses, such confounds were unlikely to account for all of the unpredictable variability of the therapists’ responses, because the five expert listeners were susceptible to the same potentially disruptive factors, but did not show similar inconsistencies.

As exemplified by the transcripts listed in Table A1 of the Appendix, the therapists who had lower intelligibility scores than predicted often transcribed target words by similar sounding pseudowords. One might speculate that they did so even in cases where they actually may have had a good guess of what was said, especially when most or all other listeners, i.e., experts and crowd, transcribed the correct word. Their decision to document the distorted phonetic surface forms rather than the corresponding lexical content may have been motivated by their awareness of the ambiguous auditory information and, consciously or not, by a desire, as committed therapists, to do diagnostic justice to the presence of the articulation problems they recognized in their clients' utterances. While the crowd- and the expert listeners may have had a competitive spirit to produce as many correct transcripts as possible, some of the therapists may have been guided by a therapeutic ambition to express their awareness of the patients' deficits. While they were actually supposed to listen for lexical content, they maintained their therapeutic role as listeners for phonetic form.

***The* n = 1 *problem in clinical assessment***

In clinical practice, diagnostic data on communication limitations are usually based on individual judgments of the treating therapists, without any further statistical verification. Clinical workflow precludes other solutions, such as involving multiple examiners or laypersons who are unfamiliar with patients (Hirsch et al., 2022). The present study replicated this limitation, - inevitably, since each of the participating PWD was supervised by only one responsible therapist. In contrast, the expert data were averaged over five individual listeners, each of whom rated all patients, providing us with a stable database against which to compare the crowd- and therapists’ responses. Even the layperson’s scores, though treated on an *n=1* basis in the present analyses, resulted from a weighted averaging of nine crowd worker scores each (Ziegler et al., 2021).

Relying on only one observation in the therapists’ assessments can cause unwanted noise in the data, as described in the previous section. Though such individual factors could not account completely for the variability of the therapists’ responses observed in the present study, they do have the potential to influence the results of structured assessments in clinical practice, over and above the expected patient-specific adaptation advantages or the detrimental effects of therapeutic biases, as described above. Taken together, these sources of heterogeneity compromise the validity of a therapist’s unique judgments of intelligibility, listening effort, or naturalness in the patients they care for.

It is worth noting that the problems associated with the n=1 condition in communication-directed assessments are less detrimental when it comes to evaluating speech motor subsystem parameters for therapeutic decision making in daily clinical work. First, the assessment of respiratory, vocal, or articulatory problems draws on SLTs' acquired skills and expertise, and there is no particular conflict or systematic bias in using this expertise to diagnose the individual patient they are caring for. Second, in clinical practice, the therapist's treatment-oriented assessment of the patient's speech breathing, voice, or articulation problems is constantly validated and increasingly enriched during the course of a therapeutic intervention, while at the same time the ability to adopt a lay perspective is constantly declining.

***Clinical consequences***

From a general perspective, communication-related parameters derived from the speech signal are necessarily limited in their ecological validity, simply because constructs such as intelligibility, listener effort, or naturalness in the “real world” are highly dependent on signal-external factors such as situational and linguistic context, environmental conditions, and the familiarity of interlocutors with the PWD. However, on the condition of reliability and reproducibility such signal-based parameters provide valuable longitudinal and cross-sectional information about a patient’s speech disorder and may allow for predictions at least within their range of validity. So, for example, parameters based on ratings from naïve, crowd-sourced listeners may be representative of laypersons’ perceptions of dysarthric speech, but less so for interlocutors in the patient's immediate environment, i.e., family or friends. As shown here, however, SLTs are neither reliable nor representative test listeners in assessments of their own clients’ communicative speech parameters, by virtue of their unique position as experts who become increasingly familiar with their clients’ speaking over treatment sessions, perceive it analytically through their trained phonetic lens, and have a therapeutic mandate towards them.

Given the time and economic constraints and the lack of access to naïve listeners in clinical care or private practices (Hirsch et al., 2022), there is no easy way out of this diagnostic dilemma. A future solution to the problem could be seen in automatic assessment approaches. Especially for intelligibility assessment, machine learning based methods are currently being investigated extensively, sometimes reporting high classification accuracies. However, such promising results so far often lack generalizability to new patients, other dysarthria types and severities, different technical set-ups and the conditions of real-life clinical diagnostics, whereas approaches based on more rigorous machine learning technologies still yield disappointing results (Gutz et al., 2022; Hall et al., 2022). Therefore, the road to automated models that successfully mimic human listeners’ perception of all varieties of dysarthric speech - from mild to severe and from flaccid to hyperkinetic - still seems to be long.

An alternative that has been proposed more recently is to recruit naïve listeners via crowdsourcing as a viable means to bring the real-world perspective to the therapist’s desk (McAllister Byun et al., 2015; Nightingale et al., 2020). The KommPaS web app that was used for data collection in the present study has been designed as a clinically usable tool that considers data security and privacy issues as well as principles of test economy, clinical time constraints and psychometric quality (Lehner, Pfab, et al., 2022). In particular, it provides sensitive and specific measures of intelligibility, listening effort and naturalness and fulfills reliability and validity criteria (Lehner, Ziegler, et al., 2022a, 2022b; Ziegler et al., 2021) and therefore applies as an instrument for daily clinical use, especially also as an outcome measure in the monitoring of disease progression or therapeutic efficiency.

# Limitations and future directions

This study has several limitations that constrain the generalizability of the results.

1. Due to the limited sample size, the multivariate correlation-based results remained exploratory and could not be substantiated by confirmatory statistical analyses. An extension of the sample by patients with very mild impairments could have strengthened the accuracy of the regression models at the upper ends of the listening effort and naturalness scales. More patients with moderate impairments could have improved the accuracy of the curvilinear regression models.
2. The patient sample could have been more balanced with regard to etiologies and syndromes. Dysarthria types differ in the regularity of their speech characteristics, which is known to have an influence on perceptual adaptation (Lansford et al., 2019).
3. The five expert listeners assessed all patients, whereas the therapists (with a few exceptions) and the crowd listener panels evaluated only a single PWD each. Hence, the expert listeners could have profited from the opportunity to compare between successively heard PWD, which may have resulted in higher consistencies.
4. The crowdsourcing stipulations of KommPaS do not prevent crowdworkers from registering for multiple listener panels, which also happened in the experiment reported here. Thus, crowd listeners can accumulate multiple exposures to dysarthric speakers. It is still unknown and needs to be examined empirically whether this leads to a measurable disorder-general adaptation effect in KommPaS.
5. The amount of professional expertise and prior experience with PWD was not controlled in the therapists. The absence of a larger patient-specific advantage might therefore be ascribable to a smaller disorder-general adaptation in therapists with less dysarthria-specific experience. However, this lack of experimental rigor also reflects the situation in real-world clinical care, where SLTs with different training backgrounds and varying levels of clinical expertise take responsibility for diagnostic assessment. Nevertheless, future studies should include parameters such as caseload of PWD, work environment, and self-perceived expertise in dealing with PWD (Borrie et al., 2021).
6. The relative contributions of implicit familiarization through frequent passive exposure to dysarthric speech vs. “professional familiarization” mediated through explicit knowledge and trained analytic listening skills on disorder-general and patient-specific adaptation cannot be discerned here, due to the lack of a suitable control group. Future studies might include occupational therapists, physiotherapists, or nurses as control listeners who are frequently exposed to dysarthric speech, but lack the expertise and auditory-perceptual skills of trained speech-language therapists.

# Conclusion

This study is the first to systematically compare speech therapists' ratings of their current clients' intelligibility, naturalness, and listening effort with the judgments of speech therapists who were unfamiliar with the patients, as well as with laypersons who had no experience with motor speech disorders and no prior knowledge of the participating patients. Our data confirmed expectations based on previous experimental studies of group- and speaker-specific adaptation effects, showing that listeners with expertise in dysarthria assessment and treatment scored overall higher than lay listeners. The response patterns of the treating therapists were specific in terms of substantial deviations from the naïve and the experienced unfamiliar listeners in both directions. Disregarding the heterogeneity in the therapists’ responses, two tendencies could be discerned: For more impaired patients, disorder-general and patient-specific adaptation advantages seemed to prevail, resulting in a tendency to substantially overestimate communicative abilities. In cases with moderate to mild impairment, on the other hand, therapeutic expertise and in-depth knowledge of their patients' dysarthria symptoms may have dominated therapists' responses and induced, consciously or not, a tendency to compensate for any adaptive advantages.

As a consequence, auditory-perceptual assessment of speech parameters by the responsible therapist, which is the gold standard in the clinical diagnostics of the motor subsystem dysfunctions in dysarthria, is fraught with validity problems when it comes to communication-relevant speech parameters. New ways must therefore be sought to find valid and reliable methods of assessing communication-related speech parameters in standard clinical care and as outcome measures of clinical trials. A future vision is that machines may one day perhaps be able to simulate human perception and comprehension of dysarthric speech. In the meantime, crowdsourcing offers a way to bring the layperson’s perspective into the therapy room.

**Acknowledgments**

This work was funded by Bayerische Sparkassenstiftung and by ReHa-Hilfe e.V. Preparatory work was supported by an ERC Proof-of-Concept grant (Proposal 737552, awarded to Jonathan Harrington) and by a PhD fellowship from the German National Academic Foundation awarded to the second author. We wish to thank all persons with dysarthria, treating therapists, clinical experts, and anonymous crowd workers for participating in this research.

**Data Availability Statement**

The conditions of our ethics approval do not permit public archiving of the data presented in this study. Readers seeking access to the data should contact the senior author [WZ] of the study. Access will be granted to named individuals in accordance with ethical procedures governing the reuse of clinical data upon request and ensuring patients’ anonymity, including completion of a formal data sharing agreement and approval of the local ethics committee.

# References

Beukelman, D. R., 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.

Borrie, S. A., Baese-Berk, M., Van Engen, K., & Bent, T. (2017). A relationship between processing speech in noise and dysarthric speech. *The Journal of the Acoustical Society of America, 141*(6), 4660-4667.

Borrie, S. A., & Lansford, K. L. (2021). A perceptual learning approach for dysarthria remediation: An updated review. *Journal of Speech, Language, and Hearing Research, 64*(8), 3060-3073.

Borrie, S. A., Lansford, K. L., & Barrett, T. S. (2021). A clinical advantage: Experience informs recognition and adaptation to a novel talker with dysarthria. *Journal of Speech, Language, and Hearing Research, 64*(5), 1503-1514.

Borrie, S. A., McAuliffe, M. J., Liss, J. M., Kirk, C., O'Beirne, G. A., & Anderson, T. (2012). Familiarisation conditions and the mechanisms that underlie improved recognition of dysarthric speech. *Language and Cognitive Processes, 27*(7-8), 1039-1055.

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., & Costantini, G. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers, 1*(1), 58. https://doi.org/10.1038/s43586-021-00055-w

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. J. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality, 54*, 13-29.

Cucchiarini, C. (1993). *Phonetic Transcription: A Methodological and Empirical Study*. CIP-Gegevens Koninklijke Bibliotheek.

D'Innocenzo, J., Tjaden, K., & Greenman, G. (2006). Intelligibility in dysarthria: Effects of listener familiarity and speaking condition. *Clinical Linguistics & Phonetics, 20*(9), 659-675.

Dagenais, P. A., Adlington, L. M., & Evans, K. J. (2011). Intelligibility, comprehensibility, and acceptability of dysarthric speech by older and younger listeners. *Journal of Medical Speech-Language Pathology, 19*(4), 37-49.

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.

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.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1975). *Motor Speech Disorders*. W.B. Saunders.

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.

Duffy, J. R. (2020). *Motor speech disorders: Substrates, differential diagnosis, and management* (4th ed.). Elsevier.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & psychophysics, 67*(2), 224-238.

Enderby, P., & Palmer, R. (2012). *Frenchay Dysarthrie Assessment - 2*. Schulz-Kirchner Verlag.

Finizia, C., Lindström, J., & Dotevall, H. (1998). Intelligibility and perceptual ratings after treatment for laryngeal cancer: laryngectomy versus radiotherapy. *The Laryngoscope, 108*(1), 138-143.

Fletcher, A. R., Wisler, A. A., Gruver, E. R., & Borrie, S. A. (2022). Beyond Speech Intelligibility: Quantifying Behavioral and Perceived Listening Effort in Response to Dysarthric Speech. *Journal of Speech, Language, and Hearing Research, 65*(11), 4060-4070.

Friedman, J., Hastie, T., & Tibshirani, R. (2019). *glasso: Graphical lasso-estimation of Gaussian graphical models.* In(Version R package version 1.11) https://CRAN.R-project.org/package=glasso

Giroud, J., Lerousseau, J. P., Pellegrino, F., & Morillon, B. (2023). The channel capacity of multilevel linguistic features constrains speech comprehension. *Cognition, 232*, 105345.

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.

Gutz, S. E., Stipancic, K. L., Yunusova, Y., Berry, J. D., & Green, J. R. (2022). Validity of off-the-shelf automatic speech recognition for assessing speech intelligibility and speech severity in speakers with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 65*(6), 2128-2143.

Hall, K., Huang, A., & Shahamiri, S. R. (2022). An Investigation to Identify Optimal Setup for Automated Assessment of Dysarthric Intelligibility using Deep Learning Technologies. *Cognitive Computation*, 1-13. https://doi.org/10.1007/s12559-022-10041-3

Hirsch, M. E., Thompson, A., Kim, Y., & Lansford, K. L. (2022). The Reliability and Validity of Speech-Language Pathologists’ Estimations of Intelligibility in Dysarthria. *Brain Sciences, 12*(8), 1011. https://doi.org/10.3390/brainsci12081011

Hustad, K. C. (2007). Effects of speech stimuli and dysarthria severity on intelligibility scores and listener confidence ratings for speakers with cerebral palsy. *Folia Phoniatrica et Logopaedica, 59*(6), 306-317.

JASP Team. (2023). *JASP.* In (Version 0.17)

Kent, R. D., & Kim, Y. (2011). The assessment of intelligibility in motor speech disorders. In A. Lowit & R. D. Kent (Eds.), *Assessment of motor speech disorders* (pp. 21-37). Plural Publishing Inc.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review, 122*(2), 148.

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.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56*(1), 1-15.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1-26.

Laganaro, M., Fougeron, C., Pernon, M., Levêque, N., Borel, S., Fournet, M., Catalano Chiuvé, S., Lopez, U., Trouville, R., & Ménard, L. (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, 35*(11), 1060-1075.

Landa, S., Pennington, L., Miller, N., Robson, S., Thompson, V., & Steen, N. (2014). Association between objective measurement of the speech intelligibility of young people with dysarthria and listener ratings of ease of understanding. *International Journal of Speech-Language Pathology, 16*(4), 408-416.

Lansford, K. L., Barrett, T. S., & Borrie, S. A. (2023). Cognitive predictors of perception and adaptation to dysarthric speech in young adult listeners. *Journal of Speech, Language, and Hearing Research, 66*(1), 1-18.

Lansford, K. L., Borrie, S. A., & Barrett, T. S. (2019). Regularity matters: Unpredictable speech degradation inhibits adaptation to dysarthric speech. *Journal of Speech, Language, and Hearing Research, 62*(12), 4282-4290.

Lehner, K., Pfab, J., & Ziegler, W. (2022). Web-based assessment of communication-related parameters in dysarthria: Development and implementation of the KommPaS web app. *Clinical Linguistics & Phonetics, 36*(12), 1093-1111.

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*, 2196-2212.

Lehner, K., Ziegler, W., & KommPaS Study Group. (2022a). Clinical measures of communication limitations in dysarthria assessed through crowdsourcing: specificity, sensitivity, and retest-reliability. *Clinical Linguistics & Phonetics, 36*(11), 988-1009.

Lehner, K., Ziegler, W., & KommPaS Study Group. (2022b). Indicators of Communication Limitation in Dysarthria and Their Relation to Auditory-Perceptual Speech Symptoms: Construct Validity of the KommPaS Web App. *Journal of Speech, Language, and Hearing Research, 65*(1), 22-42.

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*(6), 3022-3030.

Martens, H., Van Nuffelen, G., & DeBodt, M. (2010). *Nederlands Spraakverstaanbaarheidsonderzoek.* Vlaamse Vereniging voor Logopedisten.

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.

Nagle, K. F., & Eadie, T. L. (2018). Perceived listener effort as an outcome measure for disordered speech. . *Journal of communication disorders, 73*, 34-49.

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.

Pernon, M., Assal, F., Kodrasi, I., & Laganaro, M. (2022). Perceptual classification of motor speech disorders: the role of severity, speech task, and listener's expertise. *Journal of Speech, Language, and Hearing Research, 65*(8), 2727-2747.

RStudio Team. (2022). *RStudio: Integrated Development Environment for R.* In RStudio, PBC. http://www.rstudio.com/

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*(2), 216-229.

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.

Stipancic, K. L., Palmer, K. M., Rowe, H. P., Yunusova, Y., Berry, J. D., & Green, J. R. (2021). “You say severe, I say mild”: Toward an empirical classification of dysarthria severity. *Journal of Speech, Language, and Hearing Research, 64*(12), 4718-4735.

Strömbergsson, S., Edlund, J., McAllister, A., & Lagerberg, T. (2021). Understanding acceptability of disordered speech through Audience Response Systems-based evaluation. *Speech Communication, 131*, 13-22.

Tjaden, K., & Liss, J. M. (1995). The influence of familiarity on judgments of treated speech. *American journal of speech-language pathology, 4*(1), 39-48.

Whitehill, T. L., & Wong, C. C.-Y. (2006). Contributing factors to listener effort for dysarthric speech. *Journal of Medical Speech-Language Pathology, 14*(4), 335-342.

Yorkston, K. M., Beukelman, D. R., & Tice, R. (1996). *Sentence intelligibility test (SIT). Communication disorders software.* In Tice Technology Services.

Yorkston, K. M., Dowden, P. A., & Beukelman, D. R. (1992). Intelligibility measurement as a tool in the clinical management of dysarthric speakers. In R. D. Kent (Ed.), *Intelligibility in Speech Disorders. Theory, Measurement and Management* (1 ed., Vol. 1, pp. 265-285). John Benjamins Publishing Company.

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

Ziegler, W., Schölderle, T., Brendel, B., Risch, V., Felber, S., Ott, K., Goldenberg, G., Vogel, M., Bötzel, K., Zettl, L., Lorenzl, S., Lampe, R., Strecker, K., Synofzik, M., Lindig, T., Ackermann, H., & Staiger, A. (2023). Speech and nonspeech parameters in the clinical assessment of dysarthria: A dimensional analysis. *Brain Sciences, 13*, 113. https://doi.org/10.3390/brainsci13010113

Ziegler, W., Schölderle, T., Staiger, A., & Vogel, M. (2018). *Die Bogenhausener Dysarthrieskalen (BoDyS)*. Hogrefe.

Ziegler, W., & Zierdt, A. (2008). Clinical assessment of intelligibility in dysarthria: A pilot-investigation of MVP-online. *Journal of communication disorders, 41*, 553-577.

**Tables**

**Table 1.** Speakers with dysarthria: Demographic data.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Gender**  **F/M** | **Age (years)** | **Etiology** |
| 1 | M | 23 | Cerebral palsy |
| 2 | F | 55 | Cerebral palsy |
| 3 | M | 35 | Cerebral palsy |
| 4 | M | 33 | Cerebral palsy |
| 5 | M | 30 | Cerebral palsy |
| 6 | F | 43 | Cerebral palsy |
| 7 | M | 34 | Cerebral palsy |
| 8 | F | 30 | Cerebral palsy |
| 9 | F | 76 | Parkinson's disease |
| 10 | M | 77 | Stroke |
| 11 | M | 26 | Hypoxic brain injuries |
| 12 | M | 23 | Traumatic brain injury |
| 13 | M | 73 | Parkinson's disease |
| 14 | F | 53 | Stroke |
| 15 | F | 63 | Parkinson's disease |
| 16 | M | 57 | Parkinson's disease |
| 17 | M | 69 | Huntington's disease |
| 18 | F | 47 | Huntington's disease |
| 19 | M | 78 | Parkinson's disease |
| 20 | F | 71 | Amyotrophic lateral sclerosis |
| F/M = female/male | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 2:** Linear mixed modelsof *intelligibility*, *perceived listener effort* and *naturalness* by listener type, with PWD as random intercept factor. C: crowd; E: experts; T: therapists.SE: standard error,df: degrees of freedom. Intercepts are distinguished by reference categories, i.e., crowd and experts, respectively. Type III ANOVA, Satterthwaite df-method for fixed effects; LRT: likelihood ratio test statistic. Estimates of significant contrasts in bold. | | | | | |
| **Intelligibility.** *F*(2,38) = 7.69, *p* < .01; random effect: LRT(1) = 71.6, *p* < .001 | | | | | |
| **Term** | **Estimate** | **SE** | **df** | ***t*** | ***p*** |
| **Intercept (C)** | 60.30 | 5.54 | 21.47 | 10.90 | < .001 |
| **E vs. C** (disorder-general) | **8.30** | **2.36** | **38.00** | **3.53** | **< .001** |
| **T vs. C** (combined) | **7.65** | **2.36** | **38.00** | **3.25** | **< .01** |
| **Intercept (E)** | 68.60 | 5.54 | 21.47 | 12.39 | < .001 |
| **T vs. E** (patient-specific) | -0.65 | 2.36 | 38.00 | -0.28 | 0.78 |
| **Listener effort.** *F*(2,38) = 10.87, *p* < .001; random effect: LRT(1) = 38.5, *p* < .001 | | | | | |
| **Term** | **Estimate** | **SE** | **df** | ***t*** | ***p*** |
| **Intercept (C)** | 30.15 | 6.79 | 25.48 | 4.44 | < .001 |
| **E vs. C** (disorder-general) | **9.40** | **4.44** | **38.00** | **2.12** | **< .05** |
| **T vs. C** (combined) | **20.65** | **4.44** | **38.00** | **4.66** | **< .001** |
| **Intercept (E)** | 39.55 | 6.79 | 25.48 | 5.83 | < .001 |
| **T vs. E** (patient-specific) | **11.25** | **4.44** | **38.00** | **2.54** | **< .05** |
| **Naturalness.** F(2,38) =2.64, p=0.08; random effect: LRT(1) = 43.3, *p* < .001 | | | | | |
| **Term** | **Estimate** | **SE** | **df** | ***t*** | ***p*** |
| **Intercept (C)** | 35.05 | 6.22 | 24.94 | 5.64 | < .001 |
| **E vs. C** (disorder-general) | 6.55 | 3.91 | 38.00 | 1.67 | n.s. |
| **T vs. C** (combined) | **8.60** | **3.91** | **38.00** | **2.20** | **< .05** |
| **Intercept (E)** | 41.60 | 6.22 | 24.94 | 6.69 | < .001 |
| **T vs. E** (patient-specific) | 2.05 | 3.91 | 38.00 | 0.52 | n.s. |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 3:** Regression models of differences between listener types as a function of severity in terms of expert scores. **E-C:** Experts minus crowd; **T-C:** Therapists minus crowd; **T-E:** Therapists minus experts. **intc**, **lin** and **quad** indicate model coefficients, i.e. intercept, linear term and quadratic term, if applicable (empty quad-cells in linear models). Model Goodness-of-fit specified by F-statistics, R2 and residual standard error (RSE). | | | | | | | | |
| **Intelligibility.** | | | | | | | | |
| **Difference / effect** | **intc** | **lin** | **quad** | **F** | **df** | ***R2*** | ***RSE*** | ***p*** |
| **E - C** (disorder-general) | 10.87 | – 0.04 |  | 0.4 | (1,18) | 0.02 | 6.52 | n.s. |
| **T - E** (patient-specific) | 13.42 | -0.20 |  | 4.5 | (1,18) | 0.20 | 10.18 | < .05 |
| **T - C** (combined) | 24.29 | -0.24 |  | 4.6 | (1,18) | 0.21 | 11.87 | < .05 |
| **Perceived listening ease.** | | | | | | | | |
| **Difference / effect** | **intc** | **lin** | **quad** | **F** | **df** | ***R2*** | ***RSE*** | ***p*** |
| **E - C** (disorder-general) | -6.32 | 0.72 | -0.01 | 22.7 | (2,17) | 0.73 | 5.97 | < .001 |
| **T - E** (patient-specific) | 21.43 | -0.26 |  | 2.62 | (1,18) | 0.13 | 22.56 | n.s. |
| **T - C** (combined) | 12.44 | 0.69 | -0.01 | 0.62 | (2,17) | 0.07 | 23.06 | n.s. |
| **Naturalness.** | | | | | | | | |
| **Difference / effect** | **intc** | **lin** | **quad** | **F** | **df** | ***R2*** | ***RSE*** | ***p*** |
| **E - C** (disorder-general) | -12.11 | 0.92 | -0.01 | 18.9 | (2,17) | 0.69 | 6.13 | < .001 |
| **T - E** (patient-specific) | 11.77 | -0.23 |  | 2.4 | (1,18) | 0.12 | 19.26 | n.s. |
| **T - C** (combined) | -2.95 | 0.88 | -0.01 | 1.23 | (2,17) | 0.13 | 20.05 | n.s. |

**Figure captions**

**Figure 1:** Raw scores per PWD, listener and parameter. Each PWD’s score is represented by a dot. C: crowd (blue dots); E1 – E5: experts (black dots); T: therapists (red dots).

**Figure 2:** Mean values and standard errors of the three listener types and the three diagnostic parameters across PWD. Expert listener ratings E1 – E5 are averaged. For fixed effects magnitudes see Table 2.   
\* p < .05; \*\* p < .01; \*\*\* p < .001.

**Figure 3:** Adaptation as a function of severity as represented by the experts’ scores. Columns: parameters; rows: differences between listener types. E-C: experts minus crowd (disorder-general adaptation). T-E: therapists minus experts (patient-specific adaptation). T-C: therapists minus crowd (combined effect). Values > 0 indicate adaptation advantage. Solid black lines represent first or second order polynomial fits with 95% confidence intervals (shadowed areas). The dashed blue “ghost lines” in the T-C panels duplicate the regression lines of the E-C models, i.e., the disorder-general share of the combined adaptation. For model fits see Table 3. Significance levels of regression models in blue.   
\*: p < .05; \*\*\*: p < .001; n.s.: p > .05.

**Figure 4:** LASSO-regularized partial correlation network of listener responses by parameters and listener types. Edges with weights < 0.2 are suppressed. INT: intelligibility; PLE: perceived listening ease; NAT: naturalness. Extensions \_C, \_E, and \_T denote responses of the crowd-listeners, experts, and therapists, respectively.

**Figure 5:** Response consistencies of single expert listeners (black) compared with crowd listeners (blue) and therapists (red). Left: Residual standard errors describing each individual listener’s variability of regression residuals. Right: Item-total correlations per parameter.

1. We use the terms "therapists" and "experts" only to distinguish, by abbreviation, between SLTs who were currently caring for the patient they assessed and others who did not know any of the included patients. In fact, the listeners of both types were experts **and** therapists at the same time. [↑](#footnote-ref-1)
2. Clickworker has access to more than 250,000 German speaking crowd workers within an international crowd of 2.2 million. Preselection of the addressed participants is possible via a demographic filter on the platform. [↑](#footnote-ref-2)
3. Unlike the experts’ average scores, the crowd listeners‘ averages were not disassociated in this analysis. As described in the Method section, the computation of weighted averages across nine participants through an empirically established algorithm is an integral part of the crowd-sourcing method used in the KommPaS WebApp, which is necessary to protect the method from spamming (Ziegler et al., 2021). Hence, the analysis units of the KommPaS assessment consist of panels of nine listeners which cannot reasonably be broken up. [↑](#footnote-ref-3)