ASHA American Speech-Language-Hearing Association

AJSLP

Research Article

How to Efficiently Measure the Intelligibility of People With Parkinson's Disease

Kimberly L. Dahl,[a] Magdalen A. Balz,[a,b] Manuel Díaz Cádiz,[a] and Cara E. Stepp[a,b,c,d]

[a] Department of Speech, Language, and Hearing Sciences, Boston University, MA [b] Department of Gerontology, University of Massachusetts, Boston [c] Department of Biomedical Engineering, Boston University, MA [d] Department of Otolaryngology–Head and Neck Surgery, Boston University School of Medicine, MA

ARTICLE INFO

Article History:
Received February 29, 2024
Revision received June 28, 2024
Accepted August 20, 2024

Editor-in-Chief: Rita R. Patel
Editor: Jessica E. Huber

https://doi.org/10.1044/2024_AJSLP-24-00080

ABSTRACT

Purpose: The purpose of this study was to determine the most efficient approaches to measuring the intelligibility of people with Parkinson's disease (PD) when considering the estimation method, listener experience, number of listeners, number of sentences, and the ways these factors may interact.
Method: Speech-language pathologists (SLPs) and inexperienced listeners estimated the intelligibility of people with and without PD using orthographic transcription or a visual analog scale (VAS). Intelligibility estimates were based on 11 Speech Intelligibility Test sentences. We simulated all combinations of listeners and sentences to compare intelligibility estimates based on fewer listeners and sentences to a speaker-specific benchmark estimate based on the mean intelligibility across all sentences and listeners.
Results: Intelligibility estimates were closer to the benchmark (i.e., more accurate) when more listeners and sentences were included in the estimation process for transcription- and VAS-based estimates and for SLPs and inexperienced listeners. Differences between the benchmark and subset-based intelligibility estimates were, in some cases, smaller than the minimally detectable change in intelligibility for people with PD.
Conclusions: The intelligibility of people with PD can be measured more efficiently by reducing the number of listeners and/or sentences, up to a point, while maintaining the ability to detect change in this outcome. Clinicians and researchers may prioritize either fewer listeners or fewer sentences, depending on the specific constraints of their work setting. However, consideration must be given to listener experience and estimation method, as the effect of reducing the number of listeners and sentences varied with these factors.

Most people with Parkinson's disease (PD) will develop a motor speech disorder, usually in the form of hypokinetic dysarthria (Ho et al., 1999). Motor speech disorders present as a variety of articulation and voice deficits that can have a detrimental effect on intelligibility, or the degree to which the speaker's intended message is recovered by the listener (Kent et al., 1989). Reduced intelligibility is considered a core functional deficit of motor speech disorders and a key indicator of disorder severity (Stipancic et al., 2021; Weismer & Martin, 1992). This symptom can affect the speaker in substantial ways. People who are less intelligible communicate less effectively (Ball et al., 2004), encounter more interference when participating in everyday speaking situations (Borrie et al., 2022), and report reduced quality of life (Meyer et al., 2004). Measuring intelligibility is thus critical for establishing severity of impairment, monitoring treatment success, and characterizing disease progression—all common concerns in clinical practice and speech research.

Clinicians and researchers must ensure that their approach to measuring intelligibility will yield an accurate estimation of a speaker's actual intelligibility when communicating outside of the clinic. They are faced with several decisions about their approach—which estimation

Correspondence to Kimberly L. Dahl: dahl@bu.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

 Copyright © 2024 American Speech-Language-Hearing Association

Downloaded from: https://pubs.asha.org Boston University on 02/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

method to use, which listeners to recruit, and what speech sample to present. The primary goal of these decisions is, of course, to find an approach that accurately measures intelligibility. However, clinical practice and research are time- and resource-constrained endeavors, so another important goal is to find an approach that efficiently measures intelligibility while maintaining an adequate level of accuracy.

When deciding which estimation method to use, clinicians and researchers may consider orthographic transcription or a visual analog scale (VAS). Orthographic transcription entails comparing a listener's written documentation of what they heard a speaker say to the speaker's intended message. Transcription is generally considered the most objective method of measuring intelligibility. This method is often the approach taken in standardized tools, such as the Assessment of Intelligibility of Dysarthric Speech (AIDS; Yorkston & Beukelman, 1984) and its software-based counterpart, the Speech Intelligibility Test (SIT; Yorkston et al., 1996). Though its relative objectivity is appealing, orthographic transcription can be time-consuming—requiring both the time to transcribe and the time to score the transcription; it is thus infeasible in many clinical contexts. Even in well-resourced clinical practices and research settings, the time-intensiveness of orthographic transcription can be problematic. An automated scoring program is available (Borrie et al., 2019), but this does not eliminate the time for transcription. A VAS offers a faster estimation method—a listener simply marks their impression of a speaker's intelligibility on a line representing the range of possible intelligibility values. VAS ratings require no scoring or other processing. Several studies have shown that VAS-based estimates of intelligibility are moderately to strongly correlated with transcription-based estimates (Abur et al., 2019; Adams et al., 2008; Hirsch et al., 2022; Stipancic et al., 2016). The relatively fast VAS method thus offers a reasonable means of improving efficiency.

Clinicians and researchers must also decide what type of listener to recruit—experienced or inexperienced—and how many of them. Experienced listeners, such as speech-language pathologists (SLPs), represent a group that commonly evaluates intelligibility in speakers with dysarthria for diagnostic and treatment purposes. However, that experience with dysarthric speech may alter perceptions of intelligibility. Inexperienced listeners, on the other hand, may better represent the listeners whom people with dysarthria encounter in their day-to-day lives. Despite the potential effect of experience, prior work found strong relationships between intelligibility estimates from inexperienced listeners and SLPs or other trained listeners (Abur et al., 2019; Hirsch et al., 2022). This finding gives clinicians and researchers the option to turn to the type of listener more easily accessed in their work setting; experienced listeners are likely more accessible in clinical practice, and inexperienced listeners are likely more accessible in research settings.

As for the number of listeners needed, the SIT manual recommends a single listener for monitoring changes in intelligibility,[1] but later research showed that one listener never yielded an intelligibility estimate that was accurate enough for this purpose (Abur et al., 2019). Abur and colleagues did determine that the accuracy of intelligibility estimates improved as the number of listeners increased. Recruiting large groups of listeners is infeasible in clinical practice and expensive in research. Crowdsourcing may address these barriers, but evidence for the validity of this approach is preliminary (Nightingale et al., 2020), and ethical issues warrant consideration (Du et al., 2024). Fortunately, Abur and colleagues found that as few as two listeners were needed for a reasonably accurate estimate of intelligibility. This was true, however, of listeners with "extended exposure" to dysarthric speech who rated samples using a VAS; the number of listeners needed for transcription-based estimates or when the listeners are inexperienced remains unknown.

The final decision in establishing an efficient and accurate approach to measuring intelligibility involves the speech sample. No one has yet considered the impact of shorter speech samples (e.g., fewer SIT sentences) on intelligibility estimates. There is no established number of sentences needed for assessing intelligibility. The AIDS originally called for two sets of 11 sentences of increasing length in the protocol for sentence-level intelligibility. However, the second set was found to be unnecessary and potentially fatiguing, so the SIT was revised to include a single set of 11 sentences of five to 15 words (Yorkston et al., 1996). This change offered a substantial gain in efficiency. It is possible, though untested, that even fewer sentences would still yield an accurate estimate of intelligibility.

The purpose of this study was therefore to determine the most efficient approaches to measuring the intelligibility of people with PD when considering the estimation method, listener experience, number of listeners, number of sentences, and the ways these factors may interact. To address these aims in a way that is relevant to both clinical practice and research, we recruited two listener types—SLPs and listeners inexperienced in assessing dysarthric speech—and collected intelligibility estimates using two common methods—orthographic transcription and a VAS. All listeners assessed the intelligibility of 11 SIT sentences read aloud by 20 speakers with PD and four control speakers. We compared intelligibility estimates based on

[1]Multiple listeners are recommended "to establish functional levels or to compare individuals" (Yorkston et al., 1996, p. 8).

Downloaded from: https://pubs.asha.org Boston University on 02/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Table 1. Demographic and disease characteristics of speakers with Parkinson's disease (PD) and control speakers.

Speaker	Group	Age (years)	Sex[a]	Gender[a]	Dysarthria severity[b]	MDS-UPDRS III: motor severity	Years since PD diagnosis
P01[c]	PD	60	F	NR	3	41	4
P02[c]	PD	69	F	NR	21	48	1
P03[d]	PD	54	F	Woman	25	36	6
P04	PD	68	F	NR	28	2	5
P05[c]	PD	68	F	NR	30	42	14
P06	PD	68	F	NR	38	51	6
P07[d]	PD	65	F	Woman	39	48	10
P08[d]	PD	73	F	Woman	45	60	9
P09[c]	PD	77	F	NR	57	85	16
P10[c]	PD	68	F	NR	70	40	11
P11[d]	PD	68	M	Man	8	38	6
P12	PD	60	M	NR	14	36	3
P13[d]	PD	73	M	Man	27	63	4
P14[d]	PD	73	M	Man	34	22	7
P15[c]	PD	70	M	NR	35	16	4
P16[d]	PD	68	M	NR	46	20	4
P17	PD	62	M	Man	47	47	12
P18	PD	63	M	NR	51	53	6
P19[d]	PD	67	M	Man	68	77	11
P20[c]	PD	59	M	NR	76	47	5
C01	Control	61	F	Woman	—	—	—
C02	Control	68	F	Woman	—	—	—
C03	Control	61	M	Man	—	—	—
C04	Control	67	M	Man	—	—	—

Note. Participants are sorted by dysarthria severity within each group and sex. MDS-UPDRS III = Movement Disorder Society revision of the Unified Parkinson's Disease Rating Scale, Part III (0–132); F = female; NR = not recorded at the time of data collection; M = male; — = not applicable.

[a]Sex and gender were self-reported by participants. [b]Dysarthria severity was characterized as a certified speech-language pathologist's rating of overall severity on a 100-unit visual analog scale, with 0 = *no impairment* and 100 = *severely dysarthric*. [c]Included in the work of Abur et al. (2019). [d]Included in the work of Abur et al. (2021).

subsets of the 11 sentences and subsets of the available listeners to a speaker-specific benchmark estimate for each estimation method and listener type. The results of this study will guide clinicians and researchers in the best practices for efficiently obtaining accurate estimates of the intelligibility of people with PD.

Method

Speakers

Speakers were selected from existing databases of speech recordings[2] at Boston University and the University of Washington to include a sex-balanced[3] sample of individuals across a wide range of severity of dysarthria and motor symptoms. The sample (see Table 1) included 20 people with PD with a mean age of 66.6 years (SD = 5.6 years, range: 54–77 years) and four speakers without PD (M = 64.3 years, SD = 3.8 years, range: 61–68 years). Participants with PD had been diagnosed, on average, 7.2 years before their speech was recorded (SD = 3.9 years, range: 1–16 years). We selected people with PD because (a) a large database of recordings was available and (b) the heterogeneous and progressive nature of speech impairment in PD ensured a sample representative of various severity types. Control speakers were included toward the latter aim as well; they ensured that speakers with no speech impairment were also included in the sample. All speakers provided informed consent in accordance with the institutional review board of Boston University (#2625) or the University of Washington (#36181).

Disease severity was characterized by independently rating motor symptoms and dysarthria. Motor symptom severity ranged from mild to severe (Martínez-Martín

[2]Recordings were collected primarily for studies of acoustics and behavioral assays of neural control. Some samples, however, were also assessed for intelligibility in prior work, as noted in Table 1.
[3]Gender information is not available for all speakers.

Downloaded from: https://pubs.asha.org Boston University on 02/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

et al., 2015), as determined by scores on Part III of the Movement Disorder Society (MDS) revision of the Unified Parkinson's Disease Rating Scale (UPDRS; Goetz et al., 2008) completed by an MDS-certified rater (M = 44.9, SD = 17.6, range: 16–85). Dysarthria severity also spanned a wide range (M = 38.1, SD = 19.9, range: 3–76), as determined by ratings of two sentences of "The Rainbow Passage" (Fairbanks, 1960) on a 100-unit VAS (0 = *no dysarthria*, 100 = *severely dysarthric*) by a certified SLP blinded to participant diagnosis and study purpose. The MDS-UPDRS III and dysarthria ratings served not to establish ground truth regarding the level of motor and speech impairment of the speakers but rather to ensure our sample covered a wide range of impairment according to raters experienced in assessing the motor function and speech of people with PD.

Experienced Listeners

Two groups of SLPs experienced in dysarthria assessment but unfamiliar with the study's speakers participated. SLPs with at least 3 years of experience evaluating intelligibility in adults were eligible to participate, and they were recruited by convenience and snowball sampling approaches through the authors' professional networks. The first group of five SLPs (gender: all women; sex: all assigned female; M = 40.9 years, SD = 11.3 years, range: 27–55 years) completed an orthographic transcription task. They had an average of 14.5 years of experience (SD = 8.6 years, range: 3.5–24.0 years). The second group of 10 SLPs (gender: all women; sex: all assigned female; M = 36.4 years, SD = 9.3 years, range: 28–57 years) completed a VAS task and had, on average, 10.7 years of experience (SD = 8.4 years, range: 4–30 years). Unbalanced groups of listeners were recruited for each task because the relative objectivity of the transcription task was expected to result in less variable estimates across listeners; the greater variability expected with VAS ratings would require more listeners to achieve the same measurement accuracy. Listeners reported no relevant history of neurological, communication, or hearing disorders and provided informed consent in accordance with the Boston University Institutional Review Board.

Inexperienced Listeners

Two groups of listeners with no experience in dysarthria assessment, all unfamiliar with the speakers, were also enrolled in the study. The first group of inexperienced listeners completed the transcription task. In this group were five adults (gender: two women, two men, and one nonbinary/genderqueer individual; sex: three assigned female, two assigned male) with an average age of 21.2 years (SD = 1.3 years, range: 20–23 years). The second group completed the VAS task and included 10 adults (gender: eight women, two men; sex: eight assigned female, two assigned male) with an average age of 24.6 years (SD = 7.1 years, range: 19–43 years). Again, listeners in both groups reported no relevant history of neurological, communication, or hearing disorders. All listeners provided informed consent.

Speech Recordings

Speech recordings were collected in either a sound-treated booth with an omnidirectional condenser earset microphone (model MX153; Shure) and QuadMic II preamplifier or in a quiet room with a headset microphone (Models WH20, WH20XLR, SM35XLR; Shure) and handheld digital recorder (Model LS-10; Olympus). All recordings were sampled at 44.1 kHz with 16-bit resolution and collected with the microphone placed at a fixed distance of 7 cm from the mouth at a 45° angle.

Speakers were recorded while reading aloud 11 randomly generated SIT sentences (Yorkston et al., 1996). Each set of SIT sentences was drawn from the SIT database of 1,100 sentences containing five to 15 words, and each set included one sentence at each word count. The lexical characteristics of SIT sentences, such as word frequency and phonological similarity between words, are distributed such that they are unlikely to affect intelligibility estimates (Stipancic et al., 2023).

Though the database of SIT sentences is large enough to minimize repetition of sentences across sets, some sentences will occasionally be repeated across randomly generated sets. In this study, there were 17 sentences repeated in the final data set of 264 sentences (6%).

All sentence recordings were normalized for peak amplitude using a MATLAB (MathWorks) script. The normalized recordings were then mixed with multitalker babble consisting of four adult female and four adult male speakers with no speech impairment. The signal-to-noise ratio (SNR) of the mixed recordings was 0 dB, meaning the levels of the SIT sentence and the multitalker babble were equal. This SNR minimized floor and ceiling effects on intelligibility estimates during pilot testing.

Listening Tasks

Listening tasks were conducted remotely using Gorilla Experiment Builder (https://gorilla.sc), an online behavioral experiment platform. All listeners completed the study in a quiet environment on a desktop or laptop computer using wired headphones. Most participants completed the study off-site using their own computer; three completed the study on-site at Boston University. An

Downloaded from: https://pubs.asha.org Boston University on 02/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

experimenter administered consent and monitored all off-site sessions via videoconferencing.

Before beginning the listening task, listeners completed an open-source headphone screening (Milne et al., 2020) to confirm their use of suitable headphones in a sufficiently quiet environment. During the screening, listeners first set their headphone volume to a comfortable level that eliminated any effects of audibility on their intelligibility assessments; they maintained this level throughout the study. The screening then asked the listener to identify which of the three clips of white noise included a pure tone (i.e., a Huggins pitch task; Woods et al., 2017). Listeners passed the screening by identifying the correct clip in at least five out of six trials.

In both listening tasks, listeners rated 317 recordings presented in random order—264 sentences (11 sentences × 24 speakers) plus 53 (20%) that were randomly chosen to be repeated at the end of the set to calculate reliability. Note that intermixing speech samples from multiple speakers and repeating samples for reliability are departures from common clinical practice. These tactics allowed for experimental control by reducing the effect of listener familiarity (Borrie et al., 2012; D'Innocenzo et al., 2006; Tjaden & Liss, 1995a, 1995b) and establishing that the data were sufficiently reliable to justify further analysis.

Orthographic Transcription

For the orthographic transcription task, listeners played each sentence recording at least once and transcribed what they heard. They were allowed to play each recording a second time before submitting their transcription, in accordance with the standard instructions for sentence-level SIT stimuli (Yorkston & Beukelman, 1984). The transcription task lasted approximately 2 hr. Listeners were allowed to take breaks at any point during the session.

VAS

For the VAS task, listeners played each sentence and rated the speaker's intelligibility on a 100-unit VAS ranging from 0% to 100% intelligible. The VAS in the Gorilla platform appears as a continuous slider, but responses are quantized as whole numbers from 0 to 100. Intelligibility was defined for listeners as "the degree to which you understand the speech." The VAS task lasted approximately 75 min, and listeners were allowed breaks as needed.

Data Processing

VAS data required no further processing. Transcription data were manually processed by dividing the number of words correctly transcribed in a sentence by the number of total words in the target sentence and multiplying by 100. This generated a percent correct score for each sentence. Each sentence was scored independently by two researchers trained in the scoring procedures described below, which drew upon those outlined in the works of Cannito et al. (2012) and Stipancic et al. (2016). The first author reviewed any discrepancies between the two scores for a given sentence and decided on the final score for that sentence.

Words were marked correct if they were an exact match to the target, a homophone, a phonetically correct misspelling (e.g., "berch" for "birch"), a contraction error (e.g., "she's" for "she is" and vice versa), an obvious spelling error (e.g., "afriad" for "afraid"), or a reversal of consecutive words (e.g., "cover just one" for "just cover one"). Words were marked incorrect if they included an ambiguous spelling error (e.g., "both" for "booth") or errors in plurality, verb tense, or possessive marking. If a listener transcribed all words in a sentence correctly but inserted words that were not present in the target sentence (e.g., "I never wanted to be an actress" for "I wanted to be an actress"), the denominator in the percent correct calculation was increased by 1 for each inserted word.

Analysis

To evaluate the effects of the number of listeners and sentences on intelligibility estimates, we calculated for each speaker (a) a benchmark estimate of intelligibility, (b) mean intelligibility estimates based on subsets of the 11 SIT sentences and all available listeners, and (c) the absolute difference between the two. These calculations, described in more detail below, were performed separately for each listener type and estimation method.

A speaker's benchmark estimate of intelligibility was the mean intelligibility across all 11 sentences and all listeners for a given listener type and estimation method. This represented the best estimate of intelligibility for each speaker, against which other estimates were compared. We calculated four benchmark estimates for each speaker, one for each combination of listener type and estimation method (see Table 2). We standardized the benchmark calculations across estimation methods, rather than using a different approach for orthographic transcription as prescribed in the SIT (i.e., dividing the total number of correct words by the total number of words, across all sentences transcribed by all listeners).

Next, we calculated the mean intelligibility of all simulated subsets of the 11 sentences and five listeners for orthographic transcription or 10 listeners for VAS. We used a custom MATLAB script (Version 2022a) to simulate every possible combination of a speaker's intelligibility

Downloaded from: https://pubs.asha.org Boston University on 02/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Table 2. Benchmark estimates of intelligibility for each participant, by estimation method and listener type.

Speaker	Orthographic transcription		Visual analog scale	
	(a) SLP	(b) Inexperienced	(c) SLP	(d) Inexperienced
P01	44.4	49.1	55.3	43.0
P02	58.9	57.9	60.4	50.9
P03	49.2	59.3	57.5	50.7
P04	85.1	83.7	86.1	74.1
P05	80.5	82.6	79.8	71.1
P06	53.8	51.2	51.2	40.2
P07	1.3	1.0	5.5	8.5
P08	80.1	78.5	78.4	63.9
P09	77.2	78.4	74.2	63.0
P10	61.1	49.2	59.1	46.0
P11	84.6	81.8	81.0	62.5
P12	67.3	60.7	58.3	50.5
P13	62.1	64.8	67.2	52.9
P14	26.0	23.5	32.6	26.6
P15	41.9	37.6	43.4	35.9
P16	54.1	58.2	57.3	48.2
P17	10.4	11.5	11.3	11.9
P18	46.0	45.9	51.6	38.5
P19	40.2	36.3	32.2	26.1
P20	55.6	56.3	52.5	38.4
C01	83.5	86.5	86.3	73.6
C02	89.0	89.3	86.8	70.0
C03	76.7	78.5	75.8	61.0
C04	89.2	85.8	88.8	75.5

Note. Note that these values are also visualized in Figure 4. SLP = speech-language pathologist; P = person with Parkinson's disease; C = control participant.

estimates in subsets of one to 11 sentences and in subsets of one to five or one to 10 listeners. For example, we simulated combinations of VAS-based intelligibility estimates taken two sentences and three listeners at a time, until we generated every possible combination of the 6,600 total combinations at this number of listeners and sentences.[4] We then calculated the mean intelligibility estimate of each combination, subtracted that mean from the benchmark estimate for this listener type and estimation method, and took the absolute value of this difference. These absolute differences were then used to evaluate the accuracy of the intelligibility estimates based on various numbers of listeners and sentences.

The number of absolute differences calculated for each speaker varied by the number of listeners and sentences included in the combinations. They ranged from five to 4,620 for orthographic transcription and from 10 to 116,424 for VAS. The absolute difference data were then used to evaluate the performance of each number of sentences and listeners in two ways, described below.

Average Performance

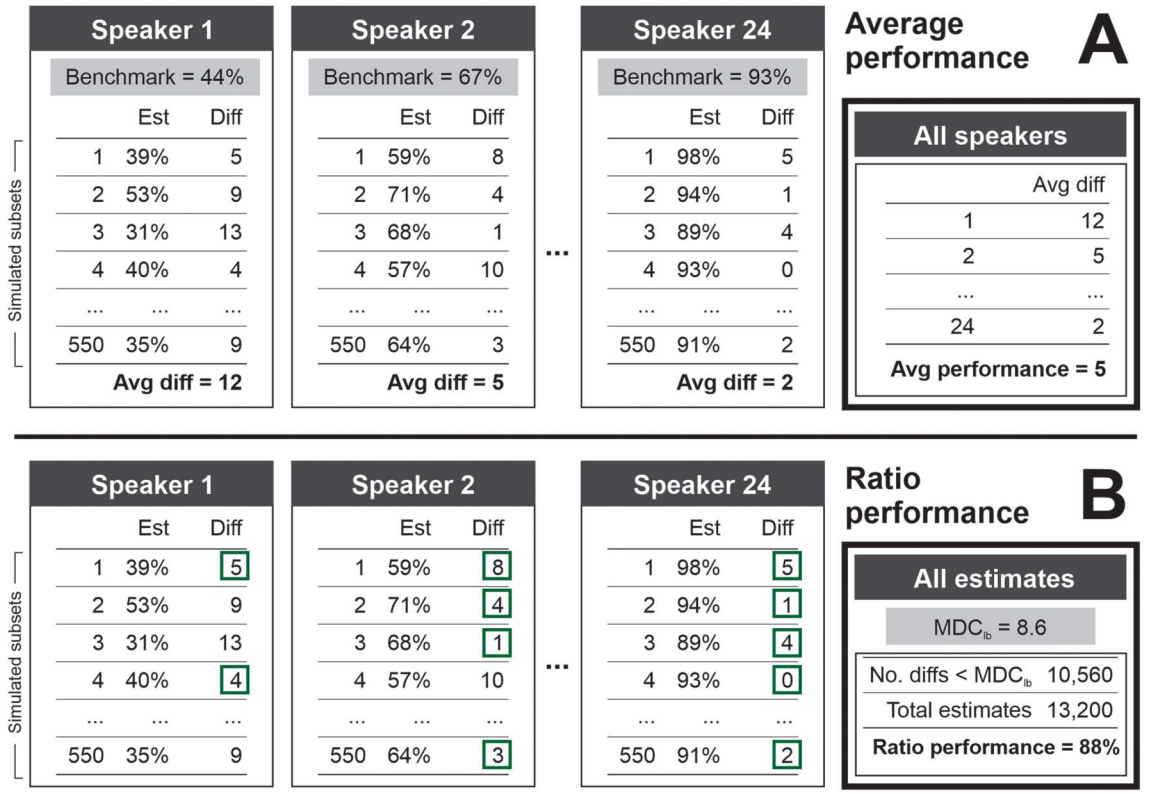
We first determined how much intelligibility estimates deviated, on average, from a speaker's benchmark when using a particular approach to measuring intelligibility. The results of this average performance analysis are useful for (a) evaluating how intelligibility estimates differ by the number of sentences and listeners involved for each estimation method and listener type and (b) comparing our findings to the larger body of research on intelligibility assessment.

To measure average performance, we first found the average deviation from the benchmark (i.e., the absolute difference described in the preceding section) for each speaker at a given number of listeners and sentences with each estimation method and listener type. We then averaged these mean values across all speakers. Figure 1A

[4]The total number of combinations of n things taken k at a time, without repetition, is calculated as $\frac{n!}{k!(n-k)!}$. Thus, the total number of combinations of two sentences (out of 11) and three listeners (out of 10 for the VAS task) is equal to $\frac{11!}{2!(11-2)!} \times \frac{10!}{3!(10-3)!}$, or 6,600 combinations.

Downloaded from: https://pubs.asha.org Boston University on 02/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Figure 1. Illustration of the process for calculating the average (Panel A) and ratio (Panel B) performance of a given approach to measuring intelligibility. In this example, these metrics are calculated for intelligibility estimates based on three speech-language pathologists transcribing nine sentences (550 estimates per speaker; 13,200 total estimates) Difference and average performance values are in percentage points; all others are percentages, as indicated. Green squares in Panel B identify intelligibility estimates that deviated from the speaker's benchmark by less than the lower bound of the minimally detectable change (MDC_{lb}) in intelligibility (Stipancic & Tjaden, 2022). Est = intelligibility estimate; diff = absolute difference between the intelligibility estimate and the speaker's benchmark; avg = average.



illustrates this process for determining the average performance of an approach to measuring intelligibility using the example of three SLPs transcribing nine sentences.

We visualized and interpreted these average performance results in the context of the minimally detectable change (MDC)[5] in intelligibility for people with PD (Stipancic & Tjaden, 2022), calculated at the 95% confidence level for intelligibility estimates from inexperienced listeners using orthographic transcription. Though this MDC was determined with a specific listener type and estimation method, we use this metric to guide our interpretation across listener types and estimation methods. We thus extended the published MDC beyond its original scope. However, with evidence of strong relationships between listener types and estimation methods (Abur et al., 2019; Adams et al., 2008; Hirsch et al., 2022; Stipancic et al., 2016) and in the absence of a published MDC specific to SLPs or VAS ratings, we considered this a justifiable application of the metric.

We will also note that Stipancic and Tjaden (2022) found the MDC in intelligibility to vary with dysarthria severity. However, in both their work and our current study, stratifying the samples by severity resulted in small subgroups. This was especially true among more impaired speakers; four speakers with severe dysarthria were included in the MDC calculation, and our sample included five. We therefore use the MDC that Stipancic and Tjaden calculated for their entire sample of participants with PD and apply it to our entire sample to ensure our interpretations are based on the most robust data available.

We used the MDC, or specifically its lower bound (MDC_{lb}), to determine if a deviation from the benchmark estimate of intelligibility was meaningful. If an estimate of intelligibility deviated from the benchmark by less than an MDC_{lb} of 8.6 percentage points, that estimate would be sufficiently accurate to detect a change in intelligibility

[5]The MDC accounts for measurement error of a given outcome by determining the "smallest magnitude of change required ... to be considered real" (Stipancic & Tjaden, 2022, p. 1858).

Downloaded from: https://pubs.asha.org Boston University on 02/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

(e.g., with treatment or disease progression) and thus still a clinically useful measure.

Though average performance provides useful information about the effect of listeners and sentences, this analysis has an important limitation—it reflects performance when the estimation process is repeated several times per speaker (up to 116,424). It thus differs notably from evaluation procedures in actual clinical practice and research. We sought to address this limitation with another analysis, described below.

Ratio Performance

In clinics and research settings, a speaker's intelligibility is not usually evaluated several times at one time point, yet such a repetitive approach—encompassing dozens, hundreds, and even thousands of estimates—is what our average performance analysis captures. To better reflect how clinicians and researchers assess intelligibility in practice and to better guide decision making in these settings, we evaluated the performance of an assessment approach if it were carried out a single time. Specifically, we calculated a measure of ratio performance that we interpreted as the likelihood that a single intelligibility estimate derived from a particular approach would be accurate enough to detect a change in intelligibility.

As above, this analysis used the data on how much an intelligibility estimate deviated from a speaker's benchmark (i.e., the absolute difference between an estimate and the benchmark). Here, we simply counted how many intelligibility estimates differed from the benchmark by less than the MDC_{lb} at a given number of listeners and sentences for each estimation method and listener type. We then divided this tally by the total number of estimates for that particular approach. This process is illustrated in Figure 1B.

To evaluate the ratio performance across different numbers of listeners and sentences, we chose a minimum acceptable ratio based on the convention in biostatistics for a minimum power ($1 - \beta$) of .80. That is, we considered a given number of listeners and sentences to offer a methodologically sound assessment approach when it yielded intelligibility estimates that deviated from the benchmark by less than the MDC_{lb} at least 80% of the time.

Relationships Between Listener Types and Estimation Methods

To provide additional context for our findings, we also evaluated the relationships between intelligibility estimates from each group of listeners and from each estimation method. We calculated two-way random intraclass correlation coefficients (ICCs) for absolute agreement of average measures (ICC(2, *k*)) to evaluate the relationship between intelligibility assessed (a) by SLPs and inexperienced listeners and (b) via orthographic transcription and VAS. Benchmark estimates of intelligibility for each speaker were used in this correlation analysis to eliminate the effects of the number of listeners and sentences on these relationships.

Reliability

Intrarater reliability was documented for each listener by calculating Pearson correlations for the 53 repeated ratings and interpreted per Cohen (1988). Intrarater reliability for the orthographic transcription task was strong for both SLPs (r_{mean} = .85, range: .78–.91) and inexperienced listeners (r_{mean} = .88, range: .81–.93). Intrarater reliability was also strong for the VAS task (SLP: r_{mean} = .78, range: .66–.83; inexperienced: r_{mean} = .85, range: .65–.89).

Interrater reliability was documented for each listener type and estimation method by calculating two-way mixed ICCs for consistency of single measures (ICC(3, 1)). ICC results were interpreted according to Koo and Li (2016). Interrater reliability for orthographic transcription was moderate, SLPs: ICC(3, 1) = .691; inexperienced: ICC(3, 1) = .734. Interrater reliability for the VAS task was excellent, SLPs: ICC(3, 1) = .939; inexperienced: ICC(3, 1) = .922.

Results

Average Performance

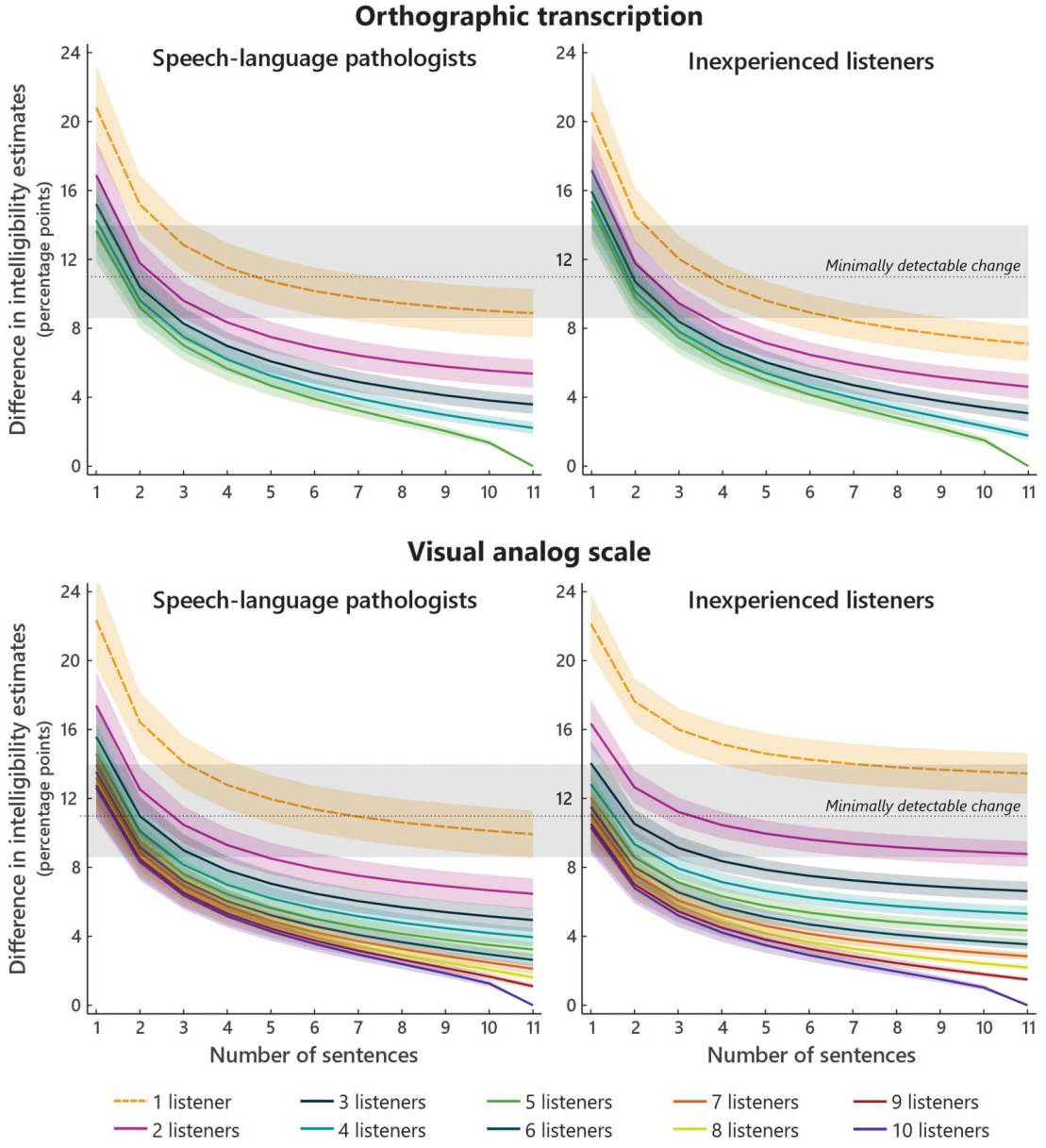
The mean absolute difference between benchmark estimates of intelligibility and estimates based on subsets of listeners and sentences ranged from 1 to 22 percentage points (see Figure 2). The overall pattern was that intelligibility estimates were closer to the benchmark (i.e., more accurate) as more listeners were included in the subset. Similarly, intelligibility estimates were closer to the benchmark as more sentences were included in the subset. This pattern held for both orthographic transcription and VAS and for both SLPs and inexperienced listeners.

Though these average performance results should not guide most clinical or research decisions, we found that, in many cases, reducing the number of listeners or sentences had a relatively small effect on average performance. The upper bound of the 95% confidence interval of the mean often remained below the MDC_{lb}, indicating that the intelligibility estimate remained sufficiently accurate, on average, to detect a change in the measure.

The fewest *listeners* that maintained this level of accuracy for orthographic transcription was one inexperienced listener transcribing at least 10 sentences, or two SLPs

Downloaded from: https://pubs.asha.org Boston University on 02/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Figure 2. Average performance: means and 95% confidence intervals of the difference between benchmark estimates of intelligibility and estimates based on simulated subsets of listeners and sentences. Dashed orange line is a single listener, with each subsequent line below it representing the inclusion of an additional listener. The horizontal dotted line and gray box across figures are the mean and range, respectively, of the minimally detectable change in intelligibility for people with Parkinson's disease (Stipancic & Tjaden, 2022).



transcribing at least five. For the VAS, accuracy was maintained with as few as two SLPs rating at least seven sentences or three inexperienced listeners rating at least five.

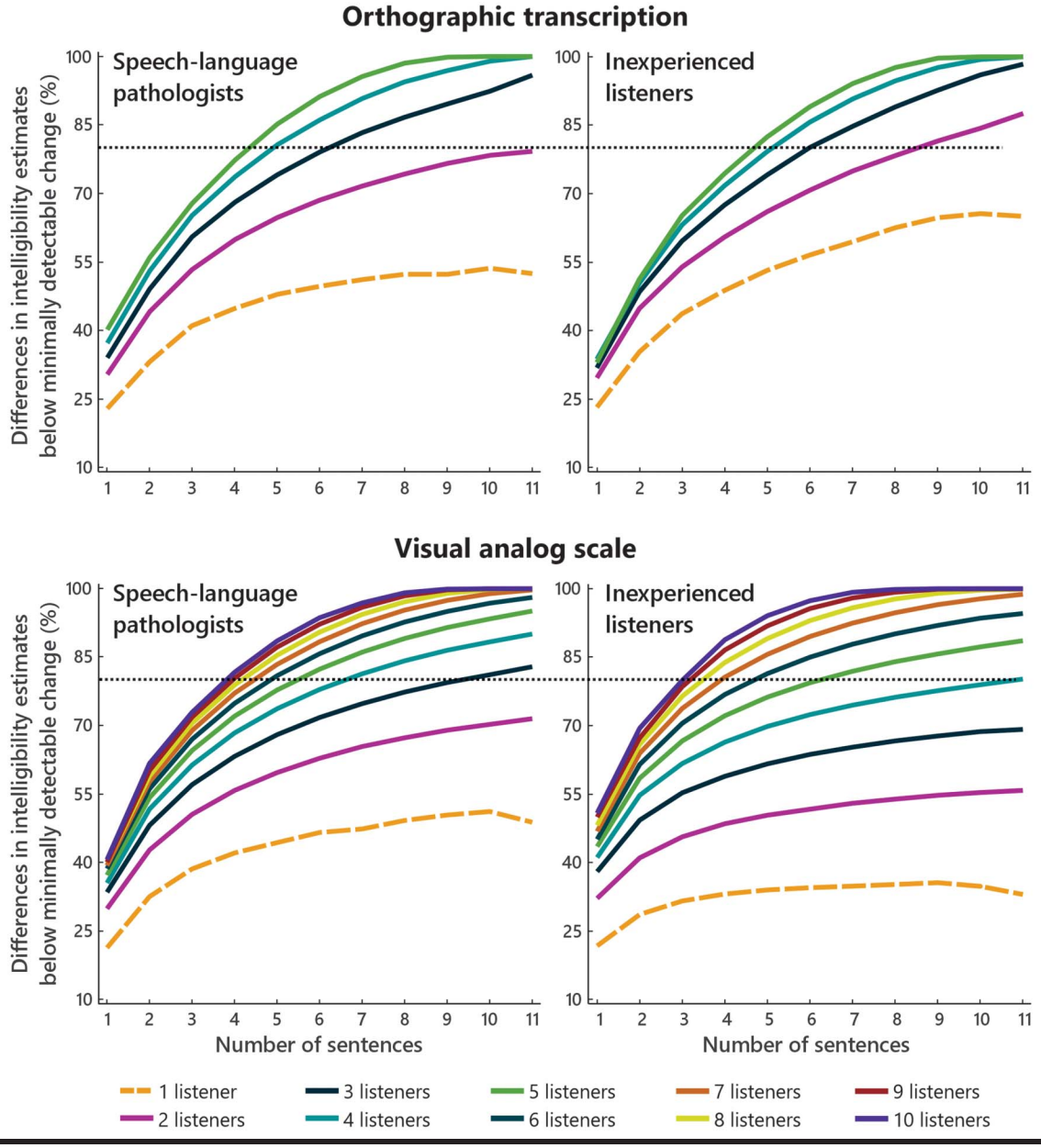
The fewest *sentences* needed to accurately measure intelligibility with orthographic transcription was three, when presented to at least four SLPs or five inexperienced listeners. The VAS required as few as three sentences, when presented to at least five SLPs, or two when presented to at least seven inexperienced listeners.

Ratio Performance

The proportion of intelligibility estimates that fell below the MDC_{lb} ranged from 21% to 100% (see Figure 3). The overall pattern was that intelligibility estimates were more likely to fall below the MDC_{lb} as more listeners and more sentences were included in the subset. This pattern again held true for both orthographic transcription and VAS and for both SLPs and inexperienced listeners.

Downloaded from: https://pubs.asha.org Boston University on 02/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Figure 3. Ratio performance: Percentage of differences in intelligibility estimates that fell below the lower bound of the minimally detectable change (MDC_{lb}) in intelligibility for people with Parkinson's disease (Stipancic & Tjaden, 2022) for each simulated subset of listeners and sentences. Dashed orange line is a single listener, with each subsequent line above it representing the inclusion of an additional listener. The horizontal dotted line marks the threshold at which 80% of intelligibility estimates were below the MDC_{lb}.



Fewer sentences or listeners often did not impede getting a sufficiently accurate estimate of intelligibility. However, the ratio performance results differed from average performance in identifying the fewest number of sentences needed to maintain sufficient accuracy.

At a minimum threshold of 80% likelihood of getting an accurate estimate, the fewest listeners needed for orthographic transcription was three SLPs transcribing at least seven sentences or two inexperienced listeners transcribing at least nine sentences. The fewest listeners needed for the VAS was three SLPs rating at least 10 sentences, or four inexperienced listeners rating all 11 sentences.

Using the same 80% threshold, the fewest sentences needed for orthographic transcription was five when presented to at least four SLPs or five inexperienced listeners. The fewest sentences needed for the VAS was four when presented to at least nine SLPs or at least seven inexperienced listeners.

Downloaded from: https://pubs.asha.org Boston University on 02/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Relationships Between Listener Types and Estimation Methods

The first relationship we tested was that of intelligibility estimates according to different listener types. There was a strong, significant relationship between estimates of intelligibility by SLPs and inexperienced listeners (see Figure 4A), for both orthographic transcription (ICC(2, k) = .992, p < .001) and VAS (ICC(2, k) = .909, p < .001). SLPs and inexperienced listeners had nearly perfect agreement with orthographic transcription. However, when using a VAS, SLPs tended to estimate higher intelligibility compared to inexperienced listeners by 10.4 percentage points on average (see Table 2, Columns c and d).

The second relationship we tested was that of intelligibility estimates from different estimation methods. There was a strong, significant relationship between estimates of intelligibility using orthographic transcription and VAS (see Figure 4B), for both SLPs (ICC(2, k) = .982, p < .001) and inexperienced listeners (ICC(2, k) = .947, p < .001). Agreement between transcription- and VAS-based estimates was nearly perfect for SLPs. However, the VAS-based estimates of inexperienced listeners were an average of 9.4 percentage points lower than transcription-based estimates from the same group (see Table 2, Columns b and d).

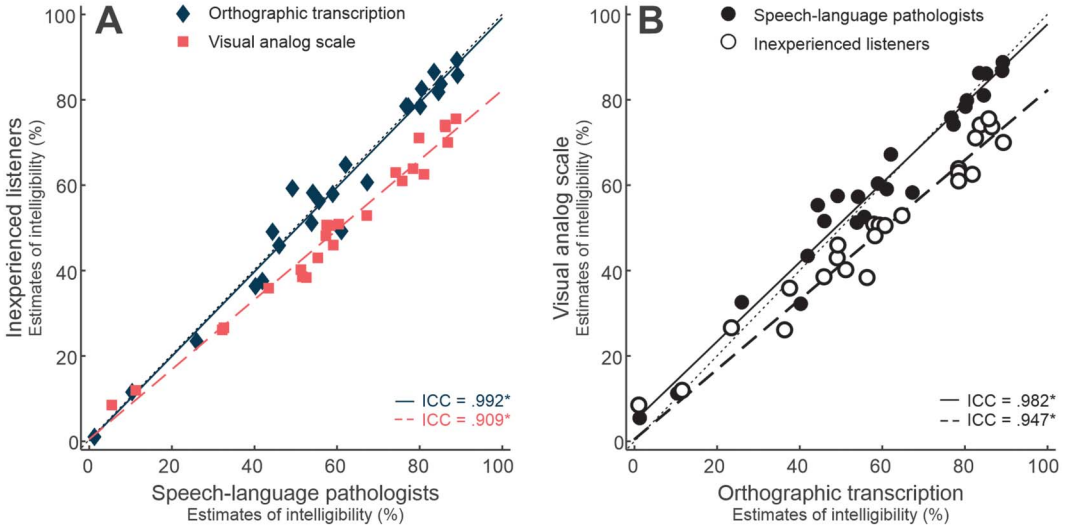
Discussion

The purpose of this study was to determine the fewest number of listeners and sentences needed across different estimation methods and listener types to efficiently measure the intelligibility of people with PD, without sacrificing measurement accuracy. We evaluated the effects of the number of listeners and sentences for both orthographic transcription and VAS and for both SLPs and inexperienced listeners. We found that, across estimation methods and listener types, intelligibility estimates were less accurate as the number of listeners and sentences was reduced, but that the drop in accuracy was sometimes small enough to justify the improvement in efficiency.

Average Effects of the Number of Listeners and Sentences

Estimates of intelligibility were, on average, closer to a speaker's benchmark estimate when more listeners and more sentences were included for both orthographic transcription and VAS and for both SLPs and inexperienced listeners. Our average performance metric showed that recruiting one fewer listener or removing a single sentence from the standard set of 11 had a small effect on the accuracy of intelligibility estimates. These efforts toward greater efficiency introduced a deviation from the benchmark by no more than 2.2 or 1.5 percentage points, respectively. However, the effect of cutting a listener or sentence grew with each subsequent removal and was, of course, cumulative. So, as additional listeners or sentences were removed, the deviation from the benchmark eventually exceeded the MDC_{lb} in intelligibility for people with PD (Stipancic & Tjaden, 2022), thus undermining the utility of the measure. The point at which this threshold of clinical utility was crossed differed by the listener type and estimation method, and the number of listeners

Figure 4. Relationships between intelligibility estimates by speech-language pathologists and inexperienced listeners (Panel A) and between intelligibility estimates using orthographic transcription and visual analog scale (Panel B). The dotted line in each panel illustrates a perfect relationship. ICC = intraclass correlation coefficient.



Downloaded from: https://pubs.asha.org Boston University on 02/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

at this crossing depended on the number of sentences and vice versa.

Using these average performance results, one might conclude that, depending on the listener type and estimation method, as few as one listener (with enough sentences) or two SIT sentences (with enough listeners) would yield an estimate of intelligibility that is sufficiently accurate to detect a change in this outcome. However, such conclusions rest on shaky ground. These results are based on the average of many repetitions of the intelligibility measurement process (up to 116,424). In actual practice, a clinician or researcher will decide on their intelligibility evaluation procedure and implement it once at a given time point. The average performance results of this study are thus a poor guide for decision making in clinical practice or research.

Improving Efficiency With Fewer Sentences or Listeners

Ratio performance, which we interpreted as the likelihood that a given estimate would be sufficiently accurate to detect a change, offers a better guide to identifying efficient approaches to intelligibility assessment. The overall trend for ratio performance was similar to that of average performance. That is, we found that the difference between an intelligibility estimate and a speaker's benchmark was more likely to fall below the MDC_{lb} when more listeners and sentences were included, for both orthographic transcription and VAS and for both SLPs and inexperienced listeners. However, the likelihood of obtaining a sufficiently accurate estimate of intelligibility with only one listener—the minimum suggested by the average performance results—ranged from 21.2% to 65.7%. Similarly, the likelihood of accurately measuring intelligibility with just two sentences ranged from 28.5% to 69.4%.

Applying our minimum threshold of 80% likelihood shows that estimating intelligibility with a single listener is never accurate enough to detect a change in intelligibility, consistent with prior work (Abur et al., 2019). The inadequacy of a single listener held even with a full set of SIT sentences and even with the relatively more objective method of orthographic transcription. A single listener will, at best, return a sufficiently accurate estimate of intelligibility only 65.7% of the time.

Abur et al. (2019) found that adding a second listener substantially improved the accuracy of intelligibility assessments. This was true of our study as well; however, the boost offered by a second listener was enough to cross the 80% threshold only for inexperienced listeners transcribing at least nine sentences. In all other cases, three or four listeners were required to yield a sufficiently accurate estimate of intelligibility at least 80% of the time.

Reaching the 80% threshold is possible through various combinations of the number of sentences and listeners, depending on the estimation method and the listener type. Recruiting fewer listeners to estimate intelligibility will require more sentences to be presented, while presenting fewer sentences will require more listeners. Clinicians and researchers can thus draw judiciously from our findings to maximize the efficiency of intelligibility evaluations by choosing to prioritize either fewer listeners or fewer sentences, given the limitations of and resources available in their work setting.

Comparing Estimation Methods and Listener Types

The overall effect of fewer listeners and sentences was qualitatively similar across estimation methods and listener types. In addition, there were strong relationships between transcription- and VAS-based estimates of intelligibility and between estimates from listeners with and without experience or training, consistent with prior research (Abur et al., 2019; Adams et al., 2008; Hirsch et al., 2022; Stipancic et al., 2016).

There were, however, some notable differences in the intelligibility estimates from the two methods and listener types included in this study. These differences offer important context for how intelligibility estimates derived from different methods and listener types should be interpreted. First, when compared to inexperienced listeners' VAS ratings, SLPs tended to estimate higher intelligibility. This finding fits with the work by Hirsch et al. (2022), who observed a similar pattern when comparing transcription-based estimates by SLPs and inexperienced listeners. Like Hirsch et al., we also found that this tendency was most pronounced for more intelligible speakers. Importantly, the average difference between estimates from SLPs and inexperienced listeners was 10.4 percentage points, a clinically meaningful difference.

Second, we found that inexperienced listeners tended to estimate lower intelligibility with VAS compared to orthographic transcription, which is typically considered the more objective approach. This was again particularly evident for more intelligible speakers. VAS-based intelligibility estimates were lower than transcription-based estimates by, on average, 9.4 percentage points, another clinically meaningful difference. Others have also documented a pattern of lower VAS-based estimates, though predominantly for speakers with severely reduced intelligibility (Abur et al., 2019). Still others found VAS to yield higher estimates than transcription, though they used alternative methods to score transcriptions that could explain the discrepancy (Adams et al., 2008; Stipancic et al., 2016). Given this range of findings, more research may be needed to fully describe the relationship between VAS- and transcription-based estimates.

Downloaded from: https://pubs.asha.org Boston University on 02/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Implications for Clinical Practice and Research

Our findings may guide clinicians and researchers seeking estimates of intelligibility that are accurate enough to detect a change in this important outcome. They show, as have others before, that estimates derived from orthographic transcription and VAS and from SLPs and inexperienced listeners are strongly related (Abur et al., 2019; Adams et al., 2008; Hirsch et al., 2022; Stipancic et al., 2016). They add to evidence specifying where clinicians and researchers can expect overestimation or underestimation of intelligibility, which may aid in interpreting results across studies or between clinical settings. Our findings add important detail to this body of work by showing how the number of listeners and sentences affect measurement accuracy for each of these estimation methods and listener types, supporting clinicians and researchers in making more informed decisions on how to efficiently measure intelligibility.

We should note, however, that although our results suggest that fewer than the full set of 11 SIT sentences may be used to assess intelligibility—with an appropriate number of listeners—they do not specify which sentence(s) should be cut. Our data cannot answer this question. There is, however, evidence that utterance length could affect intelligibility estimates in speakers with dysarthria (Allison et al., 2019; Tjaden & Wilding, 2011; Yunusova et al., 2005). The standard SIT protocol accounts for this effect by including a range of sentence lengths, so controlling for sentence length should be a factor in deciding which sentences to cut. If fewer than 11 SIT sentences will be used in intelligibility assessments, we therefore recommend that sentences are removed in a way that maintains the variation in sentence length across the sample (e.g., cutting the middle sentence or every third or fourth sentence).

Limitations

The findings of this study should be considered in the context of certain limitations. First, only dysarthria associated with PD was represented in our speaker sample. The speech and voice deficits associated with PD are heterogeneous, affecting all subsystems of speech (Ho et al., 1999), and the speaker sample covered a wide range of dysarthria severity. More importantly, intelligibility is an etiology-independent concept. It measures the degree to which a speaker's words are understood, not the reason for any misunderstandings. That is, 65% intelligibility is an equivalent outcome for a person with hypokinetic dysarthria, ataxic dysarthria, or any other form of speech impairment, even if the factors driving the reduced intelligibility (e.g., imprecise articulation, flattened prosody, atypical resonance) might differ or vary in severity. The independence of intelligibility measurement and dysarthria type or etiology is reflected in the fact that standardized assessment tools, such as the SIT, are not disorder specific and that metrics such as the MDC do not differ between, for example, people with PD and people with multiple sclerosis (Stipancic & Tjaden, 2022). We would therefore expect our findings to hold true for speakers with other types of dysarthria.

We cannot address, however, whether these findings might differ according to dysarthria severity. Stipancic and Tjaden (2022) found that the MDC in intelligibility did differ when calculated separately for groups of different severity levels, with the largest MDC for the most impaired speakers. This subgroup, however, included only four severely impaired speakers. Dividing our own sample according to dysarthria severity would create a similarly small subgroup and thus preclude meaningful analysis. Given these earlier MDC findings, however, clinicians and researchers may be justified in including more listeners and/or more sentences than the minimums suggested here when evaluating speakers with severe dysarthria.

The study stimuli are another potential limitation. Intelligibility estimates were based on SIT sentences. SLPs deploy a variety of protocols, both standardized and informal, to assess intelligibility (Gurevich & Scamihorn, 2017), and the intelligibility of speech during reading could differ from that of spontaneous speech (Kempler & Lancker, 2002). We specifically chose a reading task over more naturalistic spontaneous speech because, with the latter, we could not know with certainty what the speaker intended to say. A reading task thus lent more objectivity to the assessment. This does mean that our results may be limited to reading tasks or, more specifically, the standardized stimuli of the SIT and its paper-based predecessor, the AIDS.

Finally, the stimuli were normalized for peak amplitude and mixed with multitalker babble. These approaches are a strength of the study, as they eliminate the confound of audibility and minimize ceiling effects, respectively. However, amplitude normalization also eliminates any effect of a speaker's loudness on intelligibility, which is a common concern for people with PD and a common target of treatment. In addition, though the addition of a multitalker babble may approximate some real-world communication contexts, it does not reflect how intelligibility is usually assessed in clinical settings.

Conclusions

The accuracy of intelligibility estimates of people with PD is affected by the number of listeners evaluating the speech and the number of SIT sentences presented to

Downloaded from: https://pubs.asha.org Boston University on 02/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

those listeners. The effect of reducing the number of listeners or sentences varied according to the estimation method (orthographic transcription or VAS) and listener experience. In some cases, the effect of using fewer listeners or sentences to assess intelligibility was small enough that the resulting estimates remained sufficiently accurate for detecting changes in this outcome. We therefore conclude that the efficiency of intelligibility assessments may be improved while maintaining accuracy by recruiting fewer listeners and/or presenting fewer SIT sentences, if consideration is given to the estimation method and listener experience.

Data Availability Statement

The data sets generated during and/or analyzed during the current study are not publicly available due to commitments to protect participant confidentiality.

Acknowledgments

This work was supported by Grants DC015570 (C.E.S.), DC020867 (C.E.S. and D.D.M.), DC021080 (K.L.D.), and T32 DC013017 (C.E.S.) from the National Institute on Deafness and Other Communication Disorders and a PhD scholarship (K.L.D.) from the Council of Academic Programs in Communication Sciences and Disorders. The authors thank Rachel Norotsky and Namita Rajasubramanian for help with data analysis and Megan Cushman, Michael Madoule, Kaitlyn Siedman, and Julia Toto for help with participant recruitment.

References

Abur, D., Enos, N. M., & Stepp, C. E. (2019). Visual analog scale ratings and orthographic transcription measures of sentence intelligibility in Parkinson's disease with variable listener exposure. *American Journal of Speech-Language Pathology, 28*(3), 1222–1232. https://doi.org/10.1044/2019_AJSLP-18-0275

Abur, D., Subaciute, A., Daliri, A., Lester-Smith, R. A., Lupiani, A. A., Cilento, D., Enos, N. M., Weerathunge, H. R., Tardif, M. C., & Stepp, C. E. (2021). Feedback and feedforward auditory-motor processes for voice and articulation in Parkinson's disease. *Journal of Speech, Language, and Hearing Research, 64*(12), 4682–4694. https://doi.org/10.1044/2021_JSLHR-21-00153

Adams, S. G., Dykstra, A., Jenkins, M., & Jog, M. (2008). Speech-to-noise levels and conversational intelligibility in hypophonia and Parkinson's disease. *Journal of Medical Speech-Language Pathology, 16*(4), 165–173.

Allison, K. M., Yunusova, Y., & Green, J. R. (2019). Shorter sentence length maximizes intelligibility and speech motor performance in persons with dysarthria due to amyotrophic lateral sclerosis. *American Journal of Speech-Language Pathology, 28*(1), 96–107. https://doi.org/10.1044/2018_AJSLP-18-0049

Ball, L. J., Beukelman, D. R., & Pattee, G. L. (2004). Communication effectiveness of individuals with amyotrophic lateral sclerosis. *Journal of Communication Disorders, 37*(3), 197–215. https://doi.org/10.1016/j.jcomdis.2003.09.002

Borrie, S. A., Barrett, T. S., & Yoho, S. E. (2019). Autoscore: An open-source automated tool for scoring listener perception of speech. *The Journal of the Acoustical Society of America, 145*(1), 392–399. https://doi.org/10.1121/1.5087276

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. https://doi.org/10.1080/01690965.2011.610596

Borrie, S. A., Wynn, C. J., Berisha, V., & Barrett, T. S. (2022). From speech acoustics to communicative participation in dysarthria: Toward a causal framework. *Journal of Speech, Language, and Hearing Research, 65*(2), 405–418. https://doi.org/10.1044/2021_JSLHR-21-00306

Cannito, M. P., Suiter, D. M., Beverly, D., Chorna, L., Wolf, T., & Pfeiffer, R. M. (2012). Sentence intelligibility before and after voice treatment in speakers with idiopathic Parkinson's disease. *Journal of Voice, 26*(2), 214–219. https://doi.org/10.1016/j.jvoice.2011.08.014

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Erlbaum.

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. https://doi.org/10.1080/02699200500224272

Du, S., Babalola, M. T., D'Cruz, P., Dóci, E., Garcia-Lorenzo, L., Hassan, L., Islam, G., Newman, A., Noronha, E., & van Gils, S. (2024). The ethical, societal, and global implications of crowdsourcing research. *Journal of Business Ethics, 193,* 1–16. https://doi.org/10.1007/s10551-023-05604-9

Fairbanks, G. (1960). *Voice and articulation drillbook* (2nd ed.). Harper & Row.

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., LeWitt, P. A., Nyenhuis, D., . . . LaPelle, N. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders, 23*(15), 2129–2170. https://doi.org/10.1002/mds.22340

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

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), Article 8. https://doi.org/10.3390/brainsci12081011

Ho, A. K., Iansek, R., Marigliani, C., Bradshaw, J. L., & Gates, S. (1999). Speech impairment in a large sample of patients with Parkinson's disease. *Behavioural Neurology, 11*(3), 131–137. https://doi.org/10.1155/1999/327643

Kempler, D., & Lancker, D. V. (2002). Effect of speech task on intelligibility in dysarthria: A case study of Parkinson's disease. *Brain and Language, 80*(3), 449–464. https://doi.org/10.1006/brln.2001.2602

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders, 54*(4), 482–499. https://doi.org/10.1044/jshd.5404.482

Downloaded from: https://pubs.asha.org Boston University on 02/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

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

Martínez-Martín, P., Rodríguez-Blázquez, C., Mario Alvarez, Arakaki, T., Arillo, V. C., Chaná, P., Fernández, W., Garretto, N., Martínez-Castrillo, J. C., Rodríguez-Violante, M., Serrano-Dueñas, M., Ballesteros, D., Rojo-Abuin, J. M., Chaudhuri, K. R., & Merello, M. (2015). Parkinson's disease severity levels and MDS-Unified Parkinson's Disease Rating Scale. *Parkinsonism & Related Disorders, 21*(1), 50–54. https://doi.org/10.1016/j.parkreldis.2014.10.026

Meyer, T. K., Kuhn, J. C., Campbell, B. H., Marbella, A. M., Myers, K. B., & Layde, P. M. (2004). Speech intelligibility and quality of life in head and neck cancer survivors. *The Laryngoscope, 114*(11), 1977–1981. https://doi.org/10.1097/01.mlg.0000147932.36885.9e

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Billig, A. J., & Chait, M. (2020). *An online headphone screening test based on dichotic pitch*. bioRxiv. https://doi.org/10.1101/2020.07.21.214395

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

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. https://doi.org/10.1044/2021_JSLHR-21-00197

Stipancic, K. L., & Tjaden, K. (2022). Minimally detectable change of speech intelligibility in speakers with multiple sclerosis and Parkinson's disease. *Journal of Speech, Language, and Hearing Research, 65*(5), 1858–1866. https://doi.org/10.1044/2022_JSLHR-21-00648

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

Stipancic, K. L., Wilding, G., & Tjaden, K. (2023). Lexical characteristics of the Speech Intelligibility Test: Effects on transcription intelligibility for speakers with multiple sclerosis and Parkinson's disease. *Journal of Speech, Language, and Hearing Research, 66*(8S), 3115–3131. https://doi.org/10.1044/2023_JSLHR-22-00279

Tjaden, K., & Liss, J. M. (1995a). The role of listener familiarity in the perception of dysarthric speech. *Clinical Linguistics & Phonetics, 9*(2), 139–154. https://doi.org/10.3109/02699209508985329

Tjaden, K., & Liss, J. M. (1995b). The influence of familiarity on judgments of treated speech. *American Journal of Speech-Language Pathology, 4*(1), 39–48. https://doi.org/10.1044/1058-0360.0401.39

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

Weismer, G., & Martin, R. E. (1992). Acoustic and perceptual approaches to the study of intelligibility. In R. D. Kent (Ed.), *Studies in speech pathology and clinical linguistics* (Vol. 1, pp. 67–118). John Benjamins. https://doi.org/10.1075/sspcl.1.04wei

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Yorkston, K. M., & Beukelman, D. R. (1984). *Assessment of Intelligibility of Dysarthric Speech*. Pro-Ed.

Yorkston, K. M., Beukelman, D. R., & Tice, R. (1996). *Speech Intelligibility Test* [Computer software]. Tice Technologies.

Yunusova, Y., Weismer, G., Kent, R. D., & Rusche, N. M. (2005). Breath-group intelligibility in dysarthria: Characteristics and underlying correlates. *Journal of Speech, Language, and Hearing Research, 48*(6), 1294–1310. https://doi.org/10.1044/1092-4388(2005/090)

Downloaded from: https://pubs.asha.org Boston University on 02/12/2025, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions