Progress toward estimating the minimal clinically important difference of intelligibility: A crowdsourced perceptual experiment

Kaila L. Stipancic, Ph.D., CCC-SLP[1], klstip@buffalo.edu

Frits van Brenk, Ph.D.[1], brenk@buffalo.edu

Mengyang Qiu, Ph.D.,[2] mengyangqiu@trentu.ca

Kris Tjaden, Ph.D., CCC-SLP[1], tjaden@buffalo.edu

[1]*Department of Communicative Disorders and Sciences, University at Buffalo,*

Buffalo, New York, USA

[2]*Department of Psychology, Trent University,*

Peterborough, Ontario, Canada

Corresponding Author: Kaila L. Stipancic
Department of Communicative Disorders and Sciences
University at Buffalo
114 Cary Hall, South Campus
Buffalo, NY 14214
Email: klstip@buffalo.edu

Conflict of Interest: The authors have no relevant conflicts of interest to disclose.

Key Words: Dysarthria, Speech, Assessment

Target Journal: *Journal of Speech, Language, and Hearing Research: Conference on Motor Speech 2024 Special Issue*

Abstract

Purpose: The purpose of the current study was to estimate the minimal clinically important difference (MCID) of sentence intelligibility in control speakers and in speakers with dysarthria due to multiple sclerosis (MS) and Parkinson's disease (PD).

Methods: Sixteen control speakers, 16 speakers with MS, and 16 speakers with PD were audio-recorded reading aloud sentences in habitual, clear, fast, loud, and slow speaking conditions. Two-hundred and forty nonexpert crowdsourced listeners heard paired conditions of the same sentence content from a speaker and indicated if one condition was more understandable than another. Listeners then used a global ratings of change scale (GROC; Jaeschke et al., 1989) to indicate *how much more understandable* that condition was than the other. Listener ratings were compared with objective intelligibility scores obtained previously (Sussman & Tjaden, 2012) via orthographic transcriptions from nonexpert listeners. Receiver operating characteristic (ROC) curves and average magnitude of intelligibility difference per level of the GROC were evaluated to determine the sensitivity, specificity, and accuracy of potential cutoff scores in intelligibility for establishing thresholds of important change.

Results: MCIDs derived from the ROC curves were invalid. However, the average magnitude of intelligibility difference derived valid and useful thresholds. The MCID of intelligibility was determined to be about 7% for a small amount of difference and about 15% for a large amount of difference.

Discussion: This work demonstrates the feasibility of the novel experimental paradigm for collecting crowdsourced perceptual data to estimate MCIDs. Results provide empirical evidence that clinical tools for the perception of intelligibility by nonexpert listeners could consist of three categories, which emerged from the data ("no difference", "a little bit of difference", "a lot of

difference"). The current work is a critical step toward development of a universal language with which to evaluate changes in intelligibility as a result of neurologic injury, disease progression, and speech-language therapy.

Introduction

There has been a recent interest in describing clinically significant changes in relevant rehabilitation outcomes, including functional speech measures. In particular, speech intelligibility, or how understandable a speaker is to a listener (Yorkston & Beukelman, 1981), is the primary goal of most speech therapy protocols for individuals with neuromotor speech disorders like dysarthria. Intelligibility is also a common speech outcome measure for monitoring decline in speech production due to neurodegenerative disease progression. Methods for evaluating speech intelligibility are well established (for examples see Abur et al., 2019; Hustad & Borrie, 2021; Miller, 2013; Stipancic et al., 2016; Sussman & Tjaden, 2012; Yorkston & Beukelman, 1981). Arguably, the gold standard for measuring speech intelligibility, as operationalized in the Speech Intelligibility Test (SIT; Yorkston et al., 2007), is for listeners to orthographically transcribe audio-recorded speech materials and subsequently compare the transcriptions to the target stimuli to obtain a percentage of words correctly transcribed (Stipancic et al., 2016). Despite the clear importance of accurate intelligibility quantification, benchmarks regarding what constitutes a real, meaningful intelligibility change are lacking. This gap in knowledge limits the ability to interpret the efficacy of therapeutic speech interventions.

In 1989, Jaeschke and colleagues were the first group of researchers to describe a concept called the minimal clinically important difference (MCID). The MCID has been defined as the smallest amount of change in an outcome measure that is perceived as relevant to a patient, a clinician, or others. Other rehabilitation disciplines have successfully defined the MCID for a multitude of clinical outcomes, such as grip strength (e.g., Bohannon, 2019), pain (e.g., Copay et al., 2018), injury and disability (e.g., Dabija & Jain, 2019), and a variety of patient-reported outcomes (e.g., Engel et al., 2018).

A necessary supplement to the MCID is the minimally detectable change (MDC) (Beninato et al., 2014; Furlan & Sterr, 2018; Riddle & Stratford, 2013; M. R. Turner et al., 2010). The MDC signals whether an observed change is outside of measurement variability/error. MDCs are often calculated using a distribution-based approach. Briefly, distribution-based approaches rely on statistical characteristics of the participant sample to determine variability in the outcome measure of interest. Although distribution-based approaches are a necessary component of defining measurement responsiveness, the MDC does not specify the *clinical relevance* of a particular change (Gatchel et al., 2010). Thus, to attain thresholds for clinically meaningful change, anchor-based approaches may be used to calculate the MCID (Gatchel et al., 2010; Hays & Woolley, 2000). Anchor-based approaches examine associations between the outcome measure of interest and an external criterion that is considered an indication of important change. Typically, the external criterion is a 'gold-standard' patient-reported outcome (more subjective anchor, e.g., quality of life) or a specified adjustment in patient management (more objective anchor e.g., health-care utilization, medication use, etc.). Anchors can also be clinician-reported outcomes or other clinical outcome tools, as appropriate. Anchor-based MCID approaches are commonly considered to reflect clinical importance (Engel et al., 2018).

Recent work sought to calculate the MDC and MCID of sentence intelligibility in individuals with dysarthria secondary to amyotrophic lateral sclerosis (ALS; Stipancic et al., 2018). The SIT (Yorkston et al., 2007) was used to determine the outcome measure of interest (i.e., intelligibility) and the ALS Functional Rating Scale-Revised (ALSFRS-R; Cedarbaum et al., 1999) was used as the external anchor scale. A total of 147 individuals with ALS and 49 controls were assessed longitudinally. The MDC of SIT-derived intelligibility was calculated using formulas standard to the rehabilitation sciences literature (see Stipancic et al., 2018). At each study visit, participants with ALS also completed the ALSFRS-R (Cedarbaum et al., 1999) which is a patient-reported outcome designed to capture patient perception of motor function. The ALSFRS-R is comprised of 12 questions about motor capacity across body regions/functions, one of which pertains to speech (i.e., "How is your speech?". The five response options range from 0 = "loss of useful speech" to 4 = "normal speech process" [Cedarbaum et al., 1999]). This speech question was employed as the external anchor for use in calculating the MCID of intelligibility. Intelligibility of participants meeting a criterion of operationally defined "true change" on the ALSFRS-R speech subscore (i.e., a change of at least one point on the speech question from one data collection session to the next) were compared to participants who experienced no change on the ALSFRS-R speech question from one data collection session to the next). Receiver operating characteristic curves (ROCs) were used to define the threshold of intelligibility change that maximized sensitivity and specificity for distinguishing 'changed' and 'unchanged' participants. Ultimately, the obtained thresholds of intelligibility change were smaller in magnitude than the calculated MDC. To reiterate, the MDC is a necessary supplement to the MCID, as it defines the smallest amount of change that is necessary for the change to be outside of measurement error, and thus, can be considered real. Therefore, by definition, the MCID must be larger than the MDC to be valid, as a cut-off for relevant change cannot be smaller than detectable change (Jacobson et al., 1999; Riddle & Stratford, 2013; Stratford & Riddle, 2012). Because the MCID calculated by Stipancic et al. (2018) for speakers with ALS was smaller than the MDC, the MCID could not be considered valid. This finding is common in the rehabilitation sciences literature (e.g., Young et al., 2009), due to limiting factors such as the lack of gold-standard anchor scales and high variability in patient/clinician-reported outcomes. The scale/outcome used to anchor MCID calculations is therefore of critical importance.

A few studies in the speech-language pathology literature have examined concepts related to the MCID. Okano and colleagues (2020) calculated MDCs and MCIDs of three patient-reported swallowing outcomes. MDCs were calculated using a distribution-based approach and MCIDs were calculated using an anchor-based approach. Resulting MCIDs were smaller than calculated MDCs, similar to Stipancic et al. (2018). Hutcheson et al (2016) estimated the MCID of the MD Anderson Dysphagia Inventory (MDADI; Chen et al., 2001) using both a distribution-based approach and an anchor-based approach. The distribution-based approach yielded an MCID (referred to as an MDC in the current work) that was smaller than the anchor-based-yielded MCID (Hutcheson et al., 2016). Therefore, this MCID can be considered useful and likely reflects a clinically relevant change in scores on the MDADI. Lastly, Marks and colleagues (2021) employed an anchor-based approach to evaluate change in a vocal effort scale for patients with vocal hyperfunction. Again, the estimated MCID was within measurement error (MDC). The authors concluded that evaluations of change in vocal effort should rely, instead, on the MDC as a threshold for clinically relevant change in the absence of a valid MCID (Marks et al., 2021). All of these studies are pertinent for establishing the need to estimate MCIDs in the field of speech pathology, and also highlight the challenges of estimating thresholds for important change.

Despite the challenges to calculating MCIDs in the speech of speech pathology, in a recent study (Stipancic, Wilding, et al., 2023), we discussed the importance of distinguishing between *statistical* significance and *clinically meaningful* significance. As an illustration, in this previous study, we found an 8% difference in intelligibility between sentences that consisted of highly frequent words from high density phonetic neighborhoods as compared to sentences comprised of less frequent words from high density neighborhoods. In our study, this 8% difference in intelligibility was not statistically significant. However, related work (Stipancic & Tjaden, 2022) suggests this 8% difference is larger than measurement error, or constitutes a *real* difference, and is *likely* clinically significant. This agrees with other authors who have suggested that an 8% intelligibility difference/change is clinically meaningful (e.g., Van Nuffelen et al., 2010). In contrast, (Rodgers et al., 2013) reported a very small sentence intelligibility difference (i.e., ~1%) on the SIT (Yorkston et al., 2007) between control speakers and speakers with multiple sclerosis (MS) that was statistically significant, but would not be considered clinically meaningful. This type of approach for evaluating outcomes (i.e., by determining clinical relevance vs. assessing statistical change alone) has been used for over a decade in the rehabilitation sciences field (Gatchel et al., 2010; McGlothlin & Lewis, 2014), but is far from common in the speech literature. Although previous work has begun to identify empirical thresholds for detectable intelligibility change (using a distribution-based approach), or change outside of measurement error (Barnett et al., 2019; Stipancic et al., 2018; Stipancic & Tjaden, 2022), the threshold for clinically meaningful change in intelligibility has not yet been established.

Using a novel experimental paradigm, the purpose of the current study was to define the MCID of sentence intelligibility for speakers with multiple sclerosis (MS) and Parkinson's disease (PD), as derived by orthographic transcriptions by nonexpert, crowdsourced listeners. MS and PD can result in perceptually dissimilar dysarthrias and are commonly associated with reduced speech intelligibility. The current study leveraged an extant database of speech materials (e.g., Stipancic et al., 2016) read in response to cues intended to modify intelligibility. We identified speakers and stimuli from the database with the aim of maximizing the range of intelligibility to enhance the likelihood of accurately defining a threshold for clinically meaningful change. The primary research question addressed was: what is the MCID of intelligibility as perceived by nonexpert listeners? The focus here was on nonexpert listeners to allow for comparison of the resulting MCIDs with our previous work (Stipancic et al., 2018; Stipancic & Tjaden, 2022) establishing MDCs of intelligibility from the transcriptions of naïve listeners. This novel paradigm could provide a framework for calculating thresholds for clinically relevant change in outcome measures across the field of speech-language pathology where they are critically needed.

Methods

The study was approved by the Institutional Review Board (IRB Protocol Number: 030-732229) through the University at Buffalo. All participants provided informed consent prior to completing study procedures.

Participants

Speakers

Speakers were recruited as part of a larger project examining the acoustic and perceptual consequences of cued speaking styles or conditions in persons diagnosed with PD and MS and control speakers. Details about speakers and procedures have previously been published (Stipancic et al., 2016; Sussman & Tjaden, 2012; Tjaden et al., 2014). The speakers and recording procedures are briefly reviewed in the following section to contextualize the current study. Forty-eight of 78 speakers in the database were selected for inclusion in the current study. The 48 speakers included 16 control speakers (i.e., speakers without MS or PD) (9 females, 7 males), 16 speakers with MS (9 females, 7 males), and 16 speakers with PD (9 females, 7 males).

Speakers were selected to 1) include an equal number of speakers across the disease groups, 2) include an equal number of females and males within each disease group, and 3) to include an even distribution of speakers with varying magnitudes of intelligibility difference across speaking conditions (see details in speech samples subsection). Table 1 displays speaker characteristics including speech intelligibility scores derived from orthographic transcriptions of the SIT completed by 42 nonexpert listeners blinded to the neurological status of the speakers (see details of this listening procedure in Sussman & Tjaden, 2012). SIT scores are provided here for the purpose of describing the overall severity of the speakers. The majority of speakers have been previously characterized as having mild dysarthria (Stipancic et al., 2016; Tjaden et al., 2014). Speakers with PD presented with perceptual characteristics consistent with hypokinetic dysarthria and speakers with MS with perceptual characteristics consistent with spastic-ataxic dysarthria.

Table 1. Demographic information of speakers.

Group	**Total *N* (females:males)**	**Age (*SD*, range)**	**SIT Intelligibility (*SD*, range)**
Control speakers	16 (9:7)	57.86 years (11.74, 27-77)	93.81% (2.24, 90.21-98.26)
Speakers with multiple sclerosis	16 (9:7)	53.19 years (11.82, 29-81)	92.92% (5.48, 78.26-97.42)
Speakers with Parkinson's disease	16 (9:7)	67.75 years (8.93, 48-78)	95.35% (10.15, 54.96-95.15)
All speakers	48 (27:21)	59.60 years (12.32, 27-81)	90.69% (7.67, 54.96-98.26)

Listeners

Two groups of listeners were employed. The first group ('transcription listeners') were nonexpert listeners whose data were collected in the lab for a previous methodological study

(Stipancic et al., 2016). Transcription listeners included 50 individuals who ranged in age from 18-29 years (*mean* = 22.38, *SD* = 2.09) and passed a hearing screening. Transcription listeners participated in person in the Motor Speech Disorders Laboratory at the University at Buffalo in Buffalo, New York. The second group of listeners ('MCID listeners') consisted of 240 prospectively recruited crowdsourced nonexpert listeners (170 female, 55 male, 9 other/prefer not to say, 5 unspecified, and 1 unknown) who ranged in age from 18-30 years (*mean* = 24.13, SD = 3.66), and were living in the United States. Table 2 displays additional demographic information for the MCID listeners. Listeners from both groups self-reported to be native speakers of American English, to have obtained a high school diploma or equivalent, to have no history of speech, language, hearing, or neurological problems, and to have no or limited experience with disordered speech. Crowdsourced participants were recruited using the crowdsourcing website Prolific (prolific.co; Palan & Schitter, 2018). Following procedures used by van Brenk et al. (2022) listeners were required to have an 80% approval rating for completed studies on Prolific and to be located in the United States. Participants were instructed to use a personal computer or laptop, as the experiment was not enabled for mobile devices or tablets. Participants were given a brief description of the experiment before reading and electronically agreeing to the IRB approved consent form. Participants were then instructed to use headphones or earphones and to sit in a quiet room while completing the experiment, after which they were asked to complete a demographic questionnaire. Participants performed a sound check by playing a sample sentence, adjusting the volume to a comfortable level, and answering a question about the sentence content. If a participant answered the question incorrectly, they were asked to re-adjust the listening volume and to try again. Participants were only allowed to continue after answering the sound check question correctly. Finally, participants practiced using the interface and experimental protocol (see below) for three speakers and speech materials from the larger database who were not identified for inclusion in the current study.

The number of crowdsourced listeners (i.e., 240) was determined based on work by McAllister Byun et al. (2015). Although the task in this earlier study differed from the task in the current study, McAllister Byun et al. (2015) found that nine crowdsourced listeners yielded results matching an "industry standard" (i.e., the modal rating across 25 experienced listeners). Therefore, to assign 10 listeners to each of the 24 lists discussed in the following sections, 240 listeners were recruited.

Table 2. Demographic information of crowdsourced listeners ('MCID listeners').

Variable	*N* (%)
Gender	
Female	170 (70.83)
Male	55 (22.92)
Other/prefer not to say	9 (3.75)
Unspecified	5 (2.08)
Unknown	1 (0.42)
Highest education level	
High school/GED	101 (42.08)
Associate degree	28 (11.58)
Bachelor degree	95 (39.60)
Master degree	14 (5.83)
Doctoral degree	2 (0.83)
Race	
American Indian/Alaska Native	3 (1.25)
Asian	25 (10.42)
Black or African American	16 (6.67)
More than one race	19 (7.92)
Other/prefer not to say	6 (2.50)
White	171 (71.25)
Ethnicity	
Hispanic or Latino	26 (10.83)
Not Hispanic or Latino	212 (88.33)
Other/prefer not to say	2 (0.83)

Location in US	
Northeast	50 (20.83)
Midwest	56 (23.33)
South	87 (36.25)
West	47 (19.58)

Procedures

Speech Samples

Speakers were recorded while reading the same 25 Harvard psychoacoustic sentences (Institute of Electrical and Electronics Engineers, 1969) in five different speaking conditions: habitual, clear, fast, loud, and slow. For the purposes of the current investigation, for each speaker, the same three sentences in each condition were chosen to present to listeners, to reduce task length and to maximize the range of intelligibility difference across the five conditions. Instructions for eliciting these conditions have been published previously (Tjaden et al., 2014). Briefly, speakers were asked to speak twice as clearly as their typical speech (clear condition), at a rate twice as fast as their typical rate (fast condition), twice as loud as their regular speaking voice (loud condition), and at a rate half as fast as their regular rate (slow condition). Speakers were recorded using an AKG C410 head-mounted microphone with a constant mouth-microphone distance, positioned 10 cm and 45° to 50° from the left oral angle. The acoustic signal was pre-amplified, low-pass filtered at 9.8 kHz, and sampled at 22 kHz. The dataset was optimized for the current study as follows. First, to maximize the opportunity to reveal clinically significant differences in intelligibility, it was desirable for some speakers to demonstrate large between-condition differences in intelligibility (e.g., between the clear and the fast condition), some speakers to demonstrate no, or very small, between-condition differences, and others to demonstrate moderate between-condition differences. By using the previously obtained transcription intelligibility scores, we examined intelligibility differences between the five conditions across the larger group of 78 speakers to identify speakers who exhibited a range of intelligibility differences between conditions. Through careful selection of a subset of speakers, between-condition intelligibility differences ranged from 0% to 65.3% across the 48 speakers.

Stimuli Preparation

The recorded stimuli for the crowdsourced listeners completing the MCID task were prepared following methods employed for the transcription task and listeners (Stipancic et al., 2016; Tjaden et al., 2014). Productions of the Harvard sentences were first normalized for peak amplitude in Goldwave (GoldWave® Inc.) to reduce differences in audibility among conditions. Because baseline intelligibility was largely preserved, as suggested by the SIT (see Table 1), sentences were mixed with 20-talker multitalker babble to achieve a signal-to-noise ratio of -3dB. This served to reduce ceiling effects and to enhance differences in intelligibility between speaking conditions. For the MCID task, the three sentences for each speaker within each condition (the same sentences for each condition for each speaker) were concatenated into a single wav file with approximately 100 ms of silence between each sentence.

Listening Task Procedure and Measures

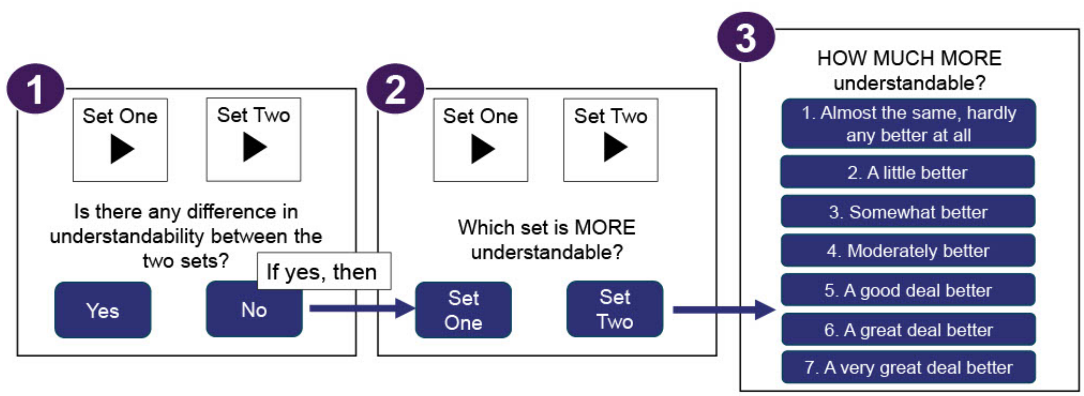
Transcription Task. The transcription task was completed in the context of a previously published study. Methodological details are available in Stipancic et al. (2016). Briefly, sentences produced by the larger cohort of 78 speakers from which the current sample of 48 speakers was chosen, were pooled and divided into 10 lists. Lists contained one sentence in each condition for each of 78 speakers. Five listeners were assigned to each list. Therefore, each sentence in each of the five conditions produced by each speaker was transcribed five times. Transcriptions were scored using a key word scoring paradigm (Hustad, 2006; Stipancic et al., 2016) in which the five key informational words (i.e., nouns, verbs, adjectives, and adverbs) in each Harvard sentence were scored as either correctly or incorrectly matching the target. The number of matches was divided by five to obtain a percentage of correctly transcribed words. For each sentence, the intelligibility scores across the five listeners were averaged to obtain an overall intelligibility score for each sentence. For the current study, scores for the three sentences of interest per condition were averaged to yield an intelligibility score for each condition. This is the intelligibility score/percentage referred to throughout the rest of this paper.

MCID Task. To control the length of the experiment for the online crowdsourced listeners, stimuli were compiled into 24 lists. First, we considered all possible two-condition comparisons (i.e., habitual-clear, habitual-fast, habitual-loud, habitual-slow, clear-fast, etc.) for an overall number of 10 condition-combinations per speaker (10 condition-combinations * 48 speakers = 480 in total). These were then divided into 24 lists following several criteria: (1) a similar number of males and females in each list; (2) a similar number of controls, speakers with MS, and speakers with PD in each list; (3) a similar number of condition-combinations (and conditions) in each list; (4) a maximum of five exposures to a given sentence within any list (to reduce the effect of familiarity with a given stimuli); and (5) never repeating a condition-combination for a given speaker within any list (to reduce the effect of familiarity with a given speaker). Each of the 24 lists contained 20 condition-combinations. On average, each list contained: (1) 10 males and 10 females (average *SD* = 2.59, range = 6-13); (2) seven speakers from each of the speaker groups (control, MS, and PD; average *SD* = 0.73, range = 5-8); (3) two of each of the possible condition-combinations (average *SD* = 0.77, range = 0-4); and (4) each of the five conditions eight times (average *SD* = 0.62, range = 4-12). In addition, within a list, no single sentence stimulus was repeated more than five times (average = 2.4, average *SD* = 1.36, range = 0-5) and no single speaker was repeated within a list more than three times (average =

1.2, *SD* = 0.09, range = 0-3). On average, the absolute difference in intelligibility between condition-combinations across lists was 13.20% (*SD* = 13.78) and ranged from 0 to 65.33%. This indicates that the magnitude of intelligibility differences, as obtained in the previous transcription study (Stipancic et al., 2016), between condition-combinations was optimized in each list as designed.

The MCID task was programmed and executed in jsPsych (De Leeuw, 2015) and hosted on Pavlovia.org (Peirce & MacAskill, 2018). Following Jaeschke et al. (1989), we used an 'external anchor of meaningfulness' in the form of a global ratings of change (GROC) scale described in the next paragraph. A visual representation of the listening task is presented in Figure 1.

Figure 1. Visual representation of crowdsourced listening task. Panel 3:Adapted from the Global Ratings of Change Scale (GROC; Jaeschke et al., 1989).



Listeners were asked to listen to the three concatenated sentences for a given speaker produced in one of the conditions ("Set One") followed immediately by the same three sentences produced in another one of the conditions ("Set Two"). Listeners were required to listen to each set completely before making their selection. They were not given a transcript or any information about what the speakers were supposed to be saying. Listeners were then asked to "Please indicate if there is any difference in understandability between the two samples" and were given response options "yes" and "no" (see panel 1 in Figure 1). If they responded "no", they moved onto the next condition. If they responded "yes", they were then asked to select which of the two samples was more understandable and chose their response by selecting "Set One" or "Set Two" (see panel 2 in Figure 1). Then, using Jaeschke's GROC scale (see panel 3 in Figure 1), they were asked "how much more understandable?" and were given response options on the seven-point scale seen in panel 3 of Figure 1. In questions one and two, stimuli sets could each be played twice. The order of speaker and condition-combination presentation were randomized across listeners by the jsPsych script.

Listeners completed this procedure for all 20 condition-combinations in their list, as well as two repeated trials interspersed for calculation of intra-rater reliability. The task took approximately 20 minutes and listeners were paid a modest fee for participating. In asking listeners to rate *understandability*, it was our intent to have listeners focus on a general concept similar to intelligibility or speech clarity, but to do so with concise and easy-to-understand terms (Weir-Mayta et al., 2017).

Thirty-nine potential listeners were excluded for failing one or more of the screening questions. The crowdsourcing website Prolific automatically excluded 72 listeners for various reasons (e.g., failing the sound check, abandoning the study prior to completion resulting in incomplete data, attempting to complete the study a second time, etc.). One participant took over the maximum allotted time of 60 minutes to complete the study, but since they answered all of the questions in the survey, we included their data. Additionally, two participants selected a large majority of “No” responses for the “Is there a difference in understandability?” question; however, since we had no reason to think this was false information, we included their data.

Data/Statistical Analyses

All statistical analyses were completed in R (Version 4.2.2, R Development Core Team, 2013).

Reliability

Reliability for the MCID task was calculated for each of the three questions displayed in Figure 1. Intrarater rater reliability was calculated for the two repeated samples that each listener responded to across the 240 listeners. Interrater reliability was calculated across the 10 listeners who heard the same list of speakers and averaged across the 24 lists. Reliability for questions 1 and 2 were calculated with Fleiss’ Kappa, and reliability for question 3 was calculated with intraclass correlation coefficients (ICC3k). All reliability analyses were completed with the *irr* package (Gamer et al., 2019). For reference, interpretation of Fleiss’ Kappa is as follows: < .00 indicates poor agreement: .00-.20 indicates slight agreement; .21-.40 indicates fair agreement; .41-.60 indicates moderate agreement; .61-.80 indicates substantial agreement; and .81-1.00 indicates almost perfect agreement (Landis & Koch, 1977). Interpretation of ICCs is as follows: < .50 indicates poor reliability; .50-.74 indicates moderate reliability; .75-.90 indicates good reliability; and > .90 indicates excellent reliability(Koo & Li, 2016).

Minimal Clinically Important Difference

Consistent with methods from studies in the rehabilitation sciences literature estimating the MCID, two analyses were conducted to calculate the MCID. These two methods involved (1) ROC curves and (2) average intelligibility difference, or a “within-patients” score difference (Copay et al., 2007). The current study followed procedures for calculating ROC curves outlined by Beninato et al. (2014) and Tilson et al. (2010) (for a similar approach see Stipancic et al. 2018). The ten condition-combinations for each of the 48 speakers (480 comparisons) were divided into groups that received ratings corresponding to each value on the GROC scale (i.e., a group of condition-combinations that were scored as being “Almost the same (1)”, a group of condition-combinations that were scored as being “A little better (2)”, etc.). Then, for each value on the GROC scale, ROC curves were calculated to determine how well the difference in percent intelligibility scores between conditions differentiated those speakers from condition-combinations for which listeners reported no difference in intelligibility (selected “No” in panel 1 of Figure 1). Each scale value of the GROC scale was examined as a potentially ‘clinically meaningful’ cut-off because, ultimately, the cut-off for what constitutes clinically meaningful change is unknown and must be empirically established. Therefore, in this initial effort to calculate MCIDs of intelligibility, it was of interest to determine which, if any, of the GROC scale values would yield valid MCID thresholds. The MCIDs were defined as the cut point from the ROC analyses that maximized both sensitivity and specificity. We also calculated the area under the curve (AUC) to identify the probability that intelligibility could distinguish between condition-combinations that listeners identified as having different understandability (i.e., for each value on the GROC scale) and condition-combinations that listeners identified as not being different in understandability. AUCs close to 0.50 indicate no better than chance probability of discriminating between speakers who had a meaningful difference in intelligibility between conditions and speakers who did not. An AUC of 0.70 is considered acceptable and AUCs of 0.80-0.90 to be excellent (Copay et al., 2007; Hosmer & Lemeshow, 2000). Thresholds that maximize sensitivity and specificity were obtained for each of the scores on the GROC scale along with their associated AUC, sensitivity, specificity, and accuracy. ROC analyses were completed with the *pROC* package (Robin et al., 2023)

A second analysis examined the average difference in intelligibility between conditions for each of the GROC scale values. For example, the intelligibility difference for all of the condition-combinations for which listeners said one of the conditions was "Somewhat better" than the other (i.e., 3 on the GROC scale), were averaged. The sensitivity, specificity, and accuracy of the intelligibility percentage difference for each score on the GROC scale were extracted from the closest threshold obtained from the ROC analyses. Finally, a linear mixed effect (LME) model containing scores on the GROC scale as a fixed effect and speaker and condition-combination as random intercepts, was conducted to examine average intelligibility differences between scores on the GROC scale (*lmerTest* package in R; Kuznetsova et al., 2017). Model diagnostics were performed to ensure that the assumptions were met. Post hoc comparisons were completed with the Tukey method with corrections for multiple comparisons using the *emmeans* package (Lenth et al., 2024).

Results

Reliability of Crowdsourced Listeners

Table 3 reports reliability statistics for the three perceptual questions in Figure 3. Because this is a novel task in the speech perception literature, expected/acceptable reliability is unknown. Moderate to good reliability was observed for the third question (i.e., HOW MUCH MORE understandable?) with ICC3s of .55 and .75 (both $p < .001$). Reliability statistics for the first (i.e., Is there any differences in the understandability between the two sets?), and second questions (i.e., Which set is MORE understandable?) were lower (i.e., Fleiss' Kappas of .27 and .14 for question one and .46 and .36 for question two, for intra-rater and inter-rater reliability respectively). For reference, there were 3,303 "yes" responses and 1,497 "no" responses to question two. Reliability statistics are considered further in the discussion.

Table 3. Reliability of crowdsourced listeners.

Question	Type of reliability	Reliability statistic used	Reliability	*p* values
1. Is there a difference in understandability? *Yes vs. No*	Intra-rater	Fleiss' Kappa	.27	< .001
	Inter-rater	Fleiss' Kappa	Mean = .14 *SD* = .07	4 lists = n.s. 4 lists < .05 16 lists < .001
2. Which stimuli are more understandable? *One vs. Two*	Intra-rater	Fleiss' Kappa	.46	< .001
	Inter-rater	Fleiss' Kappa	Mean = .36 *SD* = .10	All < .001
3. How much more understandable? *7-point GROC scale*	Intra-rater	ICC3k	.55	< .001
	Inter-rater	ICC3k	Mean = .75 *SD* = .09	All < .001

ICC: Intraclass correlation coefficient; n.s.: Not significant; SD: standard deviation

Minimal Clinically Important Difference: ROC Curves

ROC curves for each score on the GROC scale are presented in Figure 2 and associated AUCs and thresholds in Table 4. AUCs ranged from .59 to .67, indicating poor diagnostic accuracy. For all thresholds, maximizing both sensitivity and specificity resulted in a trade-off. In other words, when sensitivity was high (e.g., .79 to .87), sensitivity was low (e.g., .29 to .43).

Figure 2. Receiver operating characteristic curves for each score on the Global Ratings of Change Scale. The straight, black, diagonal line represents no better than chance of distinguishing between condition-combinations identified as being different in understandability and those identified as not being different in understandability.

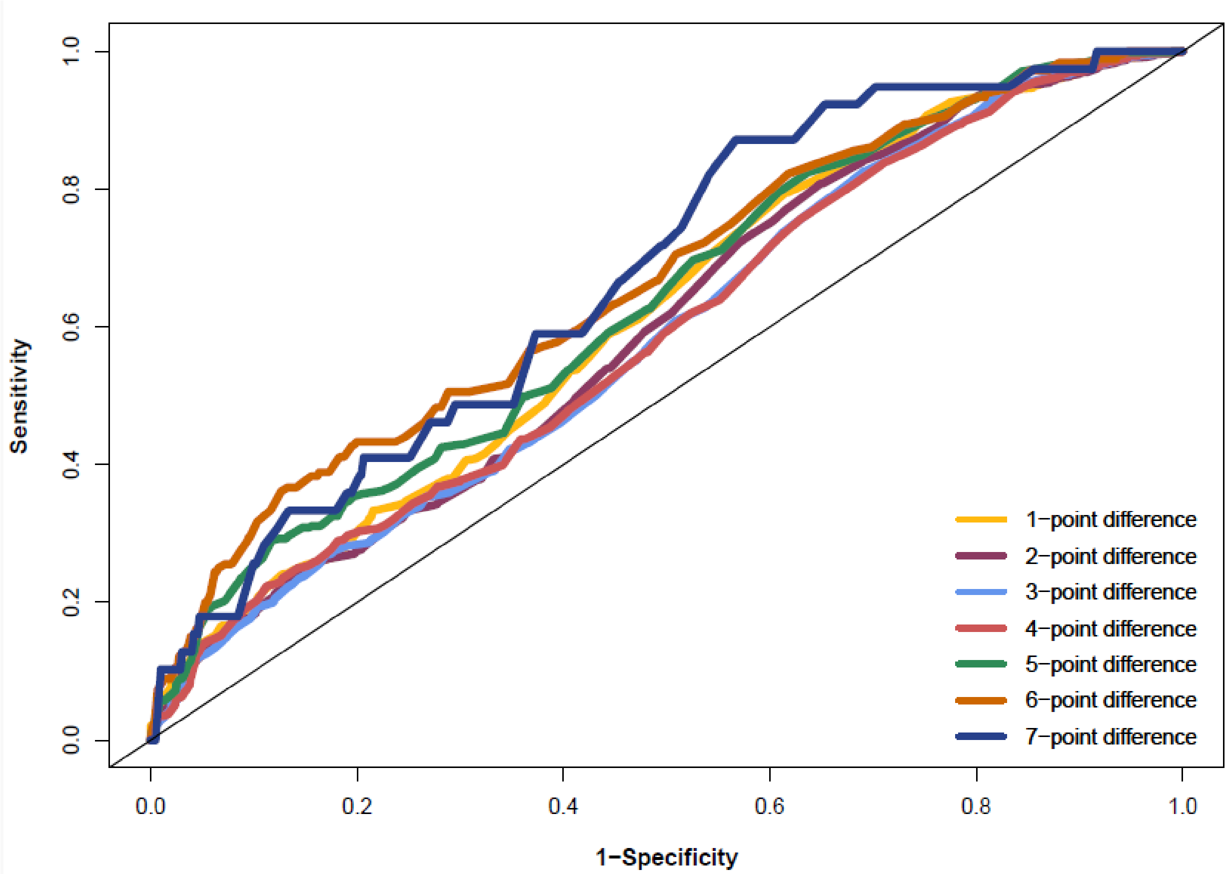


Table 4. Area under the curve and optimal thresholds for each score on the Global Ratings of Change Scale from receiver operating characteristic curve analyses.

Change on GROC Scale	*N*	AUC (95% CI)	ROC Threshold	Sensitivity	Specificity	Accuracy
1	413	.61 (.60-.63)	-3.17	0.79	0.38	0.67
2	667	.60 (.58-.61)	-3.17	0.81	0.35	0.61
3	418	.59 (.57-.60)	-3.17	0.83	0.31	0.49
4	318	.59 (.57-.61)	-3.17	0.84	0.29	0.40
5	172	.63 (.51-.66)	-1.33	0.82	0.36	0.41
6	76	.66 (.63-.70)	26.67	0.36	0.87	0.85
7	16	.67 (.59-.75)	0.67	0.87	0.43	0.44

CI: Confidence interval; GROC: Global Ratings of Change; ROC: Receiver operating characteristic

Minimal Clinically Important Difference: Average Intelligibility

Results of the LME revealed a significant main effect of GROC score, $F(1, 4755) = 32.40$, $p < .001$. Post hoc comparisons indicated significant differences between all pairs of scores ($p < .001$) except between: 1 and 2 ($p > .99$); 1 and 3 ($p = .42$); 1 and 4 ($p = .83$); 1 and 7 ($p = .10$); 2 and 3 ($p = .69$); 2 and 4 ($p = .97$); 2 and 7 ($p = .16$); 3 and 4 ($p > .99$); 3 and 7 ($p = .05$); 4 and 7 ($p = .34$); 5 and 6 ($p = .96$); 5 and 7 ($p > .99$); and 6 and 7 ($p > .99$). In summary, intelligibility scores associated with a score of 7 on the GROC scale were not statistically different from any other score on the GROC scale, potentially due to a lack of power (i.e., there were only 16 condition-combinations rated with a score of 7 on the GROC scale; see Table 4). Figure 3 displays average intelligibility differences for each score on the GROC scale. This figure illustrates three groupings of intelligibility difference scores suggested by the statistical analysis. These three groupings consisted of: (1) no difference in understandability on the GROC scale (i.e., 0); (2) a small difference in understandability on the GROC scale (i.e., 1-4; circled in purple in Figure 3); and (3) a large difference in understandability on the GROC scale (i.e., 5-7; circled in red in Figure 3).

Figure 3. Average intelligibility across the scores of the Global Ratings of Change Scale. Scores between which there was not a statistically significant difference in intelligibility are circled in purple and red (exception: score seven did not statistically differ from any other score).

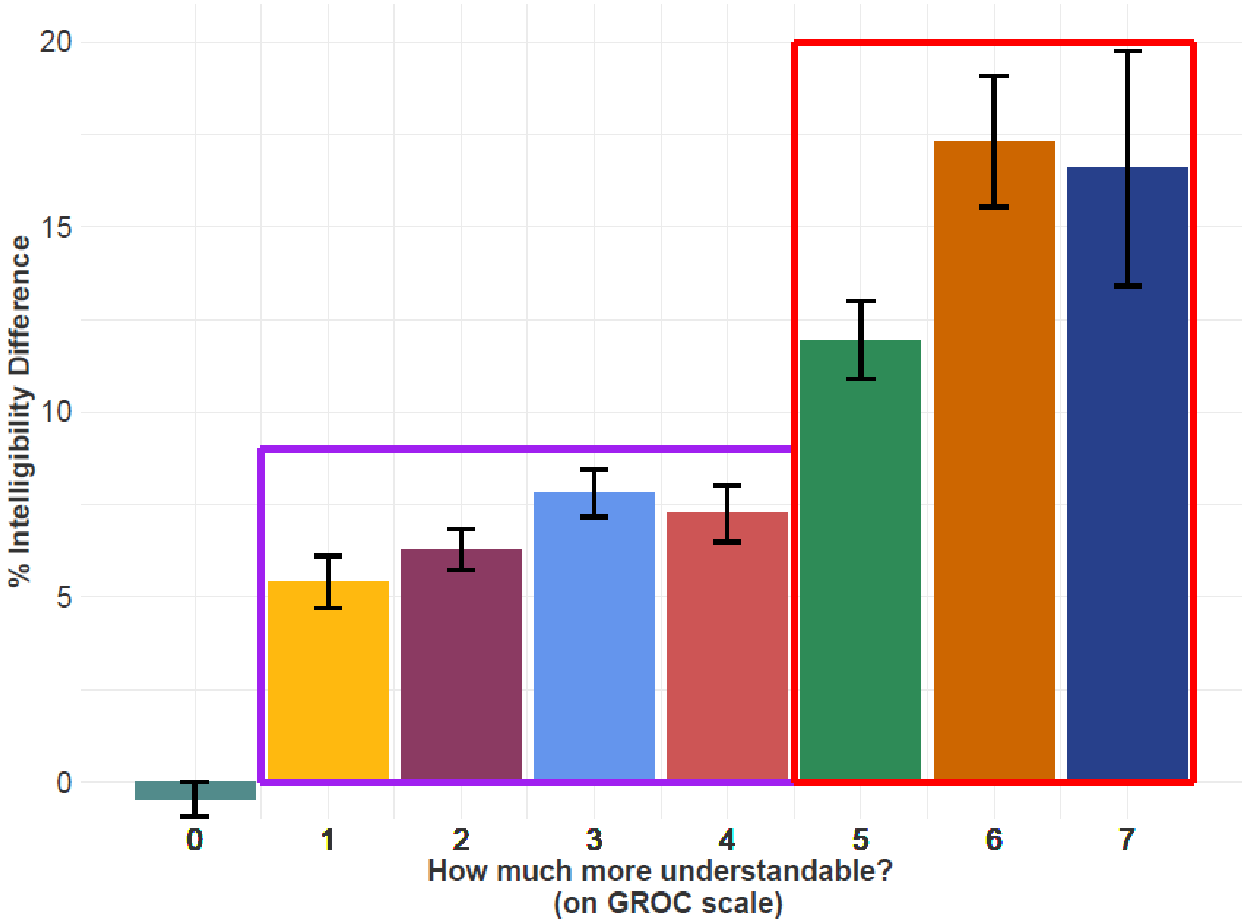


Table 5 displays the average intelligibility difference, as derived from previously obtained transcriptions in Stipancic et al. (2016), for each score on the GROC scale. The closest threshold to each of the average intelligibility differences was identified from the ROC analyses, along with the associated sensitivity, specificity, and accuracy values. All thresholds had higher specificity than sensitivity. Accuracy statistics for the higher GROC scale values (i.e., 5-7) were excellent (i.e., .71 and .79). The thresholds across adjacent GROC scale levels which were not statistically different according to the LME results were averaged to yield a single “average threshold” for three categories of MCIDs: (1) “no difference/change in intelligibility” (i.e., “no difference" between conditions per panel 1 in Figure 1); (2) "a small difference/change in intelligibility" (i.e., GROC scale scores 1-4); and (3) "a large difference/change in intelligibility" (i.e., GROC scale scores 5-7).

Table 5. Average intelligibility differences across the levels of the Global Ratings of Change Scale with sensitivity, specificity, and accuracy of the closest threshold from receiver operating characteristic curve analyses.

Difference on GROC Scale	Mean intelligibility difference % (*SD*)	Closest ROC threshold	Sensitivity	Specificity	Accuracy	Average threshold*
0 (N/A)	-0.46 (18.13)	N/A	N/A	N/A	N/A	~0%
1	5.40 (17.24)	5.33	.43	.66	.50	~7%
2	6.28 (17.47)	6.33	.42	.65	.52	
3	7.81 (17.01)	8.00	.39	.67	.57	
4	7.25 (16.96)	7.33	.40	.66	.61	
5	11.95 (18.30)	11.33	.39	.74	.71	~15%
6	17.32 (21.08)	17.33	.43	.80	.79	
7	16.58 (19.79)	16.67	.41	.79	.79	

*Average thresholds derived from averaging the mean intelligibility difference across adjacent scores on the GROC scale that were not statistically different according to the results of the linear mixed effects model.

GROC: Global Ratings of Change; ROC: Receiver operating characteristic

Discussion

This work represents the first effort to define the MCID of sentence intelligibility for speakers with dysarthria as estimated by nonexpert listeners. The current study is also one of the first in the field of speech pathology to report valid, empirically derived cut-offs of clinically meaningful change for a functional outcome measure. The thresholds provided in this work represent a meaningful advance in interpretation of intelligibility change in individuals with dysarthria and provide a framework for calculating thresholds of clinically relevant change in outcome measures across the field of speech-language pathology.

Valid MCIDs of sentence intelligibility were calculated

To reiterate, valid MCIDs must be larger in magnitude than MDCs calculated for the same population and context. Because MDCs provide a threshold for change that is outside measurement error, a threshold for clinically important change that is *within* measurement error, theoretically, cannot exist. Therefore, the MDC helps to benchmark the choice of a valid MCID. Turner et al. (2010) described how to approach such a situation: “For instance, if…two anchor-based methods (ROC and the mean change approaches) calculated on the same population yield different [MCID] values…then the knowledge that one value is below the MDC could aid in the decision to select the other” (p. 34). In the context of the current study, the MDC of intelligibility change previously calculated for mildly impaired speakers with MS and PD was, on average, 6% (Stipancic & Tjaden, 2022). In the current study, the MCIDs of intelligibility calculated with the mean change approach are larger than the previously calculated MDCs and therefore, can be considered valid. These thresholds can be further interpreted as defining a small clinically meaningful difference in intelligibility (7%) and large clinically meaningful difference (15%). These thresholds are consistent with hypotheses advanced by others, such as Van Nuffelen et al. (2010), who suggested that intelligibility changes of 8% are meaningful. Specificity and accuracy of these thresholds (obtained from the ROC analyses; see Table 5) were higher for GROC scores 5-7 than for scores 1-4. The implication is that we can have even greater confidence that an intelligibility difference closer to 15% is clinically meaningful, as compared to an intelligibility difference of approximately 7%. In addition, the threshold for a small clinically meaningful difference of 7% being close in magnitude to the previously calculated MDC of 6% should be noted. Although the MCID is larger than the MDC and thus, by definition, is a valid threshold for clinically relevant change, it should be interpreted cautiously until this result can be replicated. It is also important to consider that the previously calculated

MDC was obtained from a different context (i.e., SIT sentences, in quiet, slightly different scoring paradigm, listeners participated in person in the lab; Stipancic & Tjaden, 2022) than the MCIDs calculated here. Future studies should consider calculating MDCs and MCIDs in tandem to enhance comparability between thresholds.

In contrast, the MCIDs derived from the ROC analyses based on maximal sensitivity and specificity were not valid (see Table 4). As discussed in the introduction, this challenge of the MCID being smaller in magnitude than the MDC for the same population has arisen in previous investigations (Marks et al., 2021; Stipancic et al., 2018). Table 4 shows that the MCIDs derived from the ROC analyses were -3.17% and -1.33%, which are not only smaller than an MDC of 6%, but are also negative, which is theoretically implausible. The method for selecting the MCID threshold (i.e., at the point that maximizes both sensitivity and specificity) may have been a contributing factor. An alternative approach would be to optimize either sensitivity or specificity while sacrificing the other. However, this method would require an arbitrary decision of which threshold to select, and in the absence of any theoretical motivation to prioritize sensitivity or specificity, we followed established methods in the literature. The GROC scale value of 6 was the only GROC scale value with a valid MCID (i.e., larger than calculated MDCs), which yielded an MCID threshold of 26.67%. This finding might suggest that nonexpert listeners do not detect a clinically relevant change until there is a difference in speech that is "a great deal better" (value of 6 on GROC scale). Interestingly, ROC analyses did not yield a valid MCID for a scale value of 7 on the GROC scale (i.e., "a very great deal better"). This may be due to a variety of factors, the largest of which may be that there were only 16 condition-comparisons (see Table 4) rated as being a 7 in their difference in intelligibility. This, combined with poorer-than-ideal reliability and a large amount of variability, likely contributed to the current lack of valid results from the ROC analyses. In addition to thresholds that are smaller than MDCs, and thus, within measurement error, the AUCs for the ROC thresholds were also relatively close to 0.50, meaning that the identified thresholds are close to chance in distinguishing speakers who were identified as having a difference in intelligibility between conditions and those who were not. This calls into question the usability of such thresholds for determining clinically relevant change/difference.

The GROC scale may not be ideal for estimating MCIDs in intelligibility

Ideally, anchor-based approaches for estimating clinically important differences would rely on a gold standard functional outcome measure. Other rehabilitation science disciplines have well-established gold standard outcomes. For example, a two-point change on the Glasgow Coma Scale (Teasdale & Jennett, 1976), which is a clinician-reported outcome of neural integrity after brain injury, has been defined as a clinically important change for patients with disorders of consciousness, and thus, has been used to anchor other measures of consciousness (Mallinson et al., 2016). However, as discussed previously, such a gold standard does not currently exist for speech outcomes. Because this was the first study to investigate clinically important differences in speech intelligibility from the perspective of nonexpert listeners, using an established anchor scale (i.e., the GROC scale; Jaeschke et al., 1989) was deemed a suitable initial step. Several limitations of the GROC scale, as applied to intelligibility change, emerged. First, reliability of the GROC scale ratings (see Table 3), especially for questions one and two, were poor. However, 'adequate' reliability has not been previously established for this scale in a similar context[1]. This was a challenging perceptual task and reliability analyses removed the probability of chance agreement. Given that we averaged observations across a large number of listeners, these

[1] Reliability of a global ratings of change scale has been previously calculated for patient self-ratings of function in the physical therapy field, but not for any measures similar to the listening task described here.

statistics were deemed acceptable and provide reference values for future work using similar paradigms. Moderate to good reliability was observed for the third question.

Second, the AUCs from the ROC analyses, which are used to refer to diagnostic acceptability, were less than ideal. AUCs (see Table 4) ranged from 0.59 to 0.67, which indicates a poor diagnostic test (Carter et al., 2016). There was also a lack of statistical difference between some scores of the GROC scale for nonexpert listeners in the current study. This result may indicate that the seven-point scale gives listeners too many response options such that listeners are not able to make meaningful distinctions between adjacent scale values. Indeed, the suitability of an equal appearing interval (EAI) scale for rating intelligibility has long been criticized on psychometric grounds (Schiavetti, 1992; Schiavetti et al., 1981) such that listeners are not able to linearly partition intelligibility into equal intervals. In fact, the issue of selecting appropriate scales is still under active investigation in our field more than 30 years after Schiavetti's seminal work (Stipancic et al., in press). Therefore, the EAI of the GROC scale may not be the best way to estimate clinically important change, which was a concern levied by the scale creators. Jaeschke et al. (1989) wrote, "Despite the absence of a criterion measure, establishing the meaning of changes in a new measure requires some sort of independent standard. Global ratings represent one credible alternative" (p. 414). Future work could consider adapting the GROC scale. For example, the current results suggest that giving listeners three response options (i.e., "no difference", "a small amount of difference", "a large amount of difference") might be considered. Similarly, a study in the voice literature calculated the MCID of the Voice Handicap Index-10 (Rosen et al., 2004) by dichotomizing the anchor scale into "improvement" vs. "no improvement" in voice (V. N. Young et al., 2018). Limiting response options may also be clinically applicable for situations when a global impression of communicative function is warranted/needed for clinical decision-making. For example, having a threshold that tells us when a patient has exhibited a large amount of change (vs. no change) in intelligibility may be useful when deciding to discharge a patient from therapy.

The value of the MCID for interpreting differences/changes in speech intelligibility

The current study is an important step toward developing a universal language that researchers and clinicians can employ for interpreting intelligibility change for populations with motor speech disorders. The MCIDs provided here can be used to complement previous findings. As an example, a recent study examined the effects of a clear speech intervention for individuals with PD (Shin et al., 2022). Fifteen individuals with PD participated in eight sessions of the behavioral program. An average improvement in intelligibility of 8.53% was observed, which was determined to be statistically significant. Interestingly, the authors cite an earlier study by Beukelman et al. (2002) who also reported an 8% intelligibility improvement as a result of clear-speech-use. However, the finding in Beukelman et al. (2002) was not statistically significant, possibly owing to inadequate power (e.g., a smaller sample size) and variability across the speakers. An MCID threshold, as calculated in the current study, suggests an 8% improvement in intelligibility is still clinically meaningful (i.e., larger than the 7% threshold indicating a small meaningful difference in intelligibility). Caveats to this interpretation are discussed below.

In future studies, MCIDs of sentence intelligibility should be used to supplement statistical outcomes. MDCs have recently begun to be used in this manner (see for examples (Gutz et al., 2022; Stipancic, Golzy, et al., 2023; Stipancic, Wilding, et al., 2023). As an example of how the MCID might be deployed in the future, imagine that an intervention for individuals with MS is shown to increase intelligibility by 4%, on average, and that this magnitude of change is statistically significant. According to the MCIDs calculated in the present study, this magnitude of intelligibility change would not be considered clinically meaningful (at least to nonexpert listeners under similar conditions), and thus, the statistical significance of this finding should be interpreted with appropriate caution.

We acknowledge that findings of the current study may only apply to very similar patients in very similar contexts. Both known, and unknown, contextual effects have the potential to impact calculation of MCIDs. For example, it is important to define MCIDs for expert listeners (i.e., speech-language pathologists) to determine any effect of listener experience on estimates of clinically important differences. Therefore, when using current estimates of the MCID of intelligibility, acknowledgement of contextual factors that may differ between studies is critical. Factors such as the type of measurement (e.g., transcription, scaling, etc.), listening environment (e.g., in quiet, in background noise, etc.), stimuli characteristics (e.g., lexical and phonetic properties, amount of speech material, etc.), listener-related factors (e.g., experience, reliability, etc.), and speaker-related factors (e.g., severity, etiology, etc.) may all affect estimates of clinically important change.

Limitations & Future Directions

As highlighted by others (e.g., Gatchel et al., 2010), there is no consensus on what constitutes clinical importance, nor what external criterion should be used to anchor changes in speech outcomes. In addition, MCIDs have been found to differ based on who determines clinical importance (i.e., patients vs. clinicians) (see Beaton et al., 2002 for review). Thus, the perspective from whom significant or important changes are determined must also be considered. It should also be noted that improvements and decrements in intelligibility were considered in tandem in this study, rather than separately, as some authors have suggested (Beaton et al., 2002). As discussed by Stipancic et al. (2018), it is possible that the MCID for improvements in intelligibility may be different than the MCID for declines in intelligibility. Future work should seek to disentangle the direction of differences/changes.

Ideally, thresholds for interpreting detectable and clinically relevant change should be estimated for a wide variety of contexts for a given outcome measure (i.e., speaker group, level of speech severity, direction of change, listener type, stimuli type, listening environment, etc.). This would include calculating MCIDs separately for speakers with different etiologies of dysarthria and levels of speech impairment. For example, the MCIDs in the current study were estimated from transcriptions obtained from in-lab listeners and change scores obtained from crowdsourced listeners. Results, therefore, may have been slightly different if both groups of listeners were crowdsourced (or vice versa), as well as for different measures of intelligibility (e.g., visual analog scaling) or scoring paradigms (e.g., scoring each word in the target sentence vs. the key informational words as was done in the current study). Additionally, speech samples in the current study were presented to listeners in the presence of background noise. Although this is a valid approach and has been used in a number of previous studies from other labs (e.g., Abur et al., 2019; Darling-White & Polkowitz, 2023), MCIDs obtained in quiet listening conditions may be different. Last, in the current study, listeners only heard three sentences spoken by each speaker in different speaking conditions. The speech material (including content, length, etc.), as well as the task itself (e.g., reading, repeating, spontaneous speech, etc.), itself may affect ratings and should be considered in future studies of clinically important change.

Conclusions

Overall, this study demonstrates the feasibility of employing a novel experimental paradigm for collecting crowdsourced perceptual data, as well as establishing new data analysis methods for calculating MCIDs of speech outcomes. This work provides empirical evidence that clinical tools intended to probe the perception of intelligibility by everyday listeners could have only three response levels (i.e., “no change”, “a small amount of change”, and “a large amount of change”). The MCIDs of intelligibility reported here (i.e., a small difference of approximately 7% and a large difference of approximately 15%) are a critical step toward the development of a universal language with which to evaluate changes in intelligibility as a result of speech-language therapy and disease progression.

Acknowledgements

This work would not be possible without the time, effort, and participation of all speaker and listener participants. The authors would like to thank Emmanuelle Spronk, Nathaniel Cline, Olivia Rizzo, and Kenneth Johnson for their assistance with stimuli preparation and data collection. This work was supported by NIH-NIDCD grant R01DC004689 (PI: Kris Tjaden).

Data Availability Statement

Data supporting the results in this manuscript is available for interested researchers on request from the authors.

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

Barnett, C., Green, J. R., Marzouqah, R., Stipancic, K. L., Berry, J. D., Korngut, L., Genge, A., Shoesmith, C., Briemberg, H., Abrahao, A., Kalra, S., Zinman, L., & Yunusova, Y. (2019). Reliability and validity of speech & pause measures during passage reading in ALS. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, *21*(1–2), 42–50. https://doi.org/10.1080/21678421.2019.1697888

Beaton, D. E., Boers, M., & Wells, G. A. (2002). Many faces of the minimal clinically important difference (MCID): A literature review and directions for future research: *Current Opinion in Rheumatology*, *14*(2), 109–114. https://doi.org/10.1097/00002281-200203000-00006

Beninato, M., Fernandes, A., & Plummer, L. S. (2014). Minimal clinically important difference of the functional gait assessment in older adults. *Physical Therapy*, *94*(11), 1594–1603. https://doi.org/10.2522/ptj.20130596

Beukelman, D. R., Fager, S., Ullman, C., Hanson, E., & Logemann, J. (2002). The impact of speech supplementation and clear speech on the intelligibility and speaking rate of people with traumatic brain injury. *Journal of Medical Speech-Language Pathology*, *10*(4), 237–242.

Bohannon, R. W. (2019). Minimal clinically important difference for grip strength: A systematic review. *Journal of Physical Therapy Science*, *31*(1), 75–78. https://doi.org/10.1589/jpts.31.75

Carter, J. V., Pan, J., Rai, S. N., & Galandiuk, S. (2016). ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery*, *159*(6), 1638–1645. https://doi.org/10.1016/j.surg.2015.12.029

Cedarbaum, J. M., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B., & Nakanishi, A. (1999). The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function. *Journal of the Neurological Sciences*, *169*(1–2), 13–21. https://doi.org/10.1016/S0022-510X(99)00210-5

Chen, A., Frankowski, R., Bishop-Leone, J., Hebert, T., Leyk, S., Lewin, J., & Goepfert, H. (2001). The development and validation of a dysphagia-specific quality-of-life questionnaire for patients with head and neck cancer. *Archives of Otolaryngology - Head and Neck Surgery*, *127*(7), 870–876.

Copay, A. G., Eyberg, B., Chung, A. S., Zurcher, K. S., Chutkan, N., & Spangehl, M. J. (2018). Minimum clinically important difference: Current trends in the orthopaedic literature, part II: Lower extremity: A systematic review. *Journal of Bone & Joint Surgery*, *6*(9), e2. https://doi.org/10.2106/JBJS.RVW.17.00160

Copay, A. G., Subach, B. R., Glassman, S. D., Polly, D. W., & Schuler, T. C. (2007). Understanding the minimum clinically important difference: A review of concepts and methods. *The Spine Journal*, *7*, 541–546. https://doi.org/10.1016/j.spinee.2007.01.008

Dabija, D. I., & Jain, N. B. (2019). Minimal clinically important difference of shoulder outcome measure and diagnoses: A systematic review. *American Journal of Physical Medicine & Rehabilitation*, *98*(8), 671–676. https://doi.org/10.1097/PHM.0000000000001169

Darling-White, M., & Polkowitz, R. (2023). Sentence length effects on intelligibility in two groups of older children with neurodevelopmental disroders. *American Journal of Speech-Language Pathology*, *32*(5), 2297–2310. https://doi.org/10.1044/2023_AJSLP-23-00093

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Engel, L., Beaton, D. E., & Touma, Z. (2018). Minimal clinically important difference: A review of outcome measure score interpretation. *Rheumatic Disease Clinics of North America*, *44*(2), 177–188. https://doi.org/10.1016/j.rdc.2018.01.011

Furlan, L., & Sterr, A. (2018). The applicability of standard error of measurement and minimal detectable change to motor learning research—A behavioral study. *Frontiers in Human Neuroscience*, *12*(March), 1–10. https://doi.org/10.3389/fnhum.2018.00095

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *Package "irr."*

Gatchel, R. J., Lurie, J. D., & Mayer, T. G. (2010). Minimal clinically important difference: *Spine*, *35*(19), 1739–1743. https://doi.org/10.1097/BRS.0b013e3181d3cfc9

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 wiht amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, 1–16.

Hays, R. D., & Woolley, J. M. (2000). The concept of clinically meaningful difference in health-related quality-of-life research: How meaningful is it? *Pharmacoeconomics*, *18*(5), 419–423. https://doi.org/10.2165/00019053-200018050-00001

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression. New York: Wiley, Inc, 2000.* Wiley, Inc.

Hustad, K. C. (2006). A closer look at transcription intelligibility for speakers with dysarthria: Evaluation of scoring paradigms and linguistic errors made by listeners. *American Journal of Speech-Language Pathology*, *15*(3), 268–277. https://doi.org/10.1044/1058-0360(2006/025)

Hustad, K. C., & Borrie, S. A. (2021). Intelligibility Impairment. In J. S. Damico, N. Muller, & M. J. Ball (Eds.), *The Handbook of Language and Speech Disorders* (Vol. 2, pp. 81–94). Wiley-Blackwell. https://doi.org/10.1002/9781119606987.ch4

Hutcheson, K. A., Barrow, M. P., Lisec, A., Barringer, D. A., Gries, K., & Lewin, J. S. (2016). What is a clinically relevant difference in MDADI scores between groups of head and neck cancer patients? *The Laryngoscope*, *126*(5), 1108–1113. https://doi.org/10.1002/lary.25778

Institute of Electrical and Electronics Engineers, . (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, *17*, 225–246.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, *67*(3), 300–307. https://doi.org/10.1037/0022-006X.67.3.300

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*, *10*(4), 407–415. https://doi.org/10.1016/0197-2456(89)90005-6

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lenth, R. V., Bolker, B., Buerkner, P., Gine-Vasquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2024). *Package "emmeans."*

Mallinson, T., Pape, T. L.-B., & Guernon, A. (2016). Responsiveness, minimal detectable change, and minimally clinically important differences for the disorders of consciousness scale. *The Journal Of Head Trauma Rehabilitation*, *31*(4), E43–E51. https://doi.org/10.1097/HTR.0000000000000184

Marks, K. L., Verdi, A., Toles, L. E., Stipancic, K. L., Ortiz, A. J., Hillman, R. E., & Mehta, D. D. (2021). Psychometric analysis of an ecological vocal effort scale in individuals with and without vocal hyperfunction during activities of daily living. *American Journal of Speech-Language Pathology*, *30*(6), 2589–2604. https://doi.org/10.1044/2021_ajslp-21-00111

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

McGlothlin, A. E., & Lewis, R. J. (2014). Minimal clinically important difference: Defining what really matters to patients. *JAMA*, *312*(13), 1342–1343. https://doi.org/10.1001/jama.2014.13128

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

Okano, I., Ortiz Miller, C., Salzmann, S. N., Hoshino, Y., Shue, J., Sama, A. A., Cammisa, F. P., Girardi, F. P., & Hughes, A. P. (2020). Minimum clinically important differences of the hospital for special surgery dysphagia and dysphonia inventory and other dysphagia measurements in patients undergoing ACDF. *Clinical Orthopaedics and Related Research*, *478*(10), 2309–2320. https://doi.org/10.1097/CORR.0000000000001236

Palan, S., & Schitter, C. (2018). Prolific. Ac-A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Peirce, J., & MacAskill, M. (2018). *Building experiments in PsychoPy*. Sage.

R Development Core Team. (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.

Riddle, D., & Stratford, P. (2013). *Is this change real? Interpreting patient outcomes in physical therapy*. F.A. Davis Company.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Markus, M., Siegert, S., Doering, M., & Billings, Z. (2023). *Package "pROC."*

Rodgers, J. D., Tjaden, K., Feenaughty, L., Weinstock-Guttman, B., & Benedict, R. H. B. (2013). Influence of cognitive function on speech and articulation rate in multiple sclerosis. *Journal of the International Neuropsychological Society*, *19*(2), 173–180. https://doi.org/doi:10.1017/S1355617712001166

Rosen, C. A., Lee, A. S., Osborne, J., Zullo, T., & Murry, T. (2004). Development and validation of the Voice Handicap Index-10. *The Laryngoscope*, *114*(9), 1549–1556. https://doi.org/10.1097/00005537-200409000-00009

Schiavetti, N. (1992). Scaling procedures for the measurement of speech intelligibility. In R. D. Kent (Ed.), *Intelligibility in speech disorders: Theory, measurement and management* (pp. 11–34). John Benjamins Publishing Company.

Schiavetti, N., Metz, D. E., & Sitler, R. W. (1981). Construct validity of direct magnitude estimation and interval scaling of speech intelligibility: Evidence from a study of the hearing impaired. *Journal of Speech and Hearing Research*, *24*(3), 441–445. https://doi.org/10.1044/jshr.2403.441

Shin, H., Shivabasappa, P., & Koul, R. (2022). Effect of clear speech intervention program on speech intelligibility in persons with idiopathic Parkinson’s disease: A pilot study. *International Journal of Speech-Language Pathology*, *24*, 33–41. https://doi.org/10.1080/17549507.2021.1943522

Stipancic, K. L., Golzy, M., Zhao, Y., Pinkerton, L., Rohl, A., & Kuruvilla-Dugdale, M. (2023). Improving perceptual speech ratings: The effects of auditory training on judgments of dysarthria speech. *Journal of Speech, Language, and Hearing Research*, *66*(11), 4236–4258. https://doi.org/10.1044/2023_JSLHR-23-00322

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., Whelan, B.-M., Laur, L., Zhao, Y., Rohl, A., Choi, I., & Kuruvilla-Dugdale, M. (in press). Tipping the scales: Indiscriminate use of interval scales to rate diverse dysarthric features. *Journal of Speech, Language, and Hearing Research*.

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 diease. *Journal of Speech, Language, and Hearing Research*, *66*(8S), 3115–3131. https://doi.org/10.1044/2023_JSLHR-22-00279

Stipancic, K. L., Yunusova, Y., Berry, J. D., & Green, J. R. (2018). Minimally detectable change and minimal clinically important difference of a decline in sentence intelligibility and speaking rate for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, *61*(11), 2757–2771. https://doi.org/10.1044/2018_AJSLP-17-0074

Stratford, P. W., & Riddle, D. L. (2012). When minimal detectable change exceeds a diagnostic test-based threshold change value for an outcome measure: Resolving the conflict. *Physical Therapy*, *92*(10), 1338–1347.

Sussman, J. E., & Tjaden, K. (2012). Perceptual measures of speech from individuals with Parkinson's disease and multiple sclerosis: Intelligibility and beyond. *Journal of Speech, Language, and Hearing Research*, *55*(August), 1208–1219. https://doi.org/10.1044/1092-4388(2011/11-0048)1208

Teasdale, G., & Jennett, B. (1976). Assessment and prognosis of coma after head injury. *Acta Neurochirurgic*, *34*(1–4), 45–55. https://doi.org/10.1007/BF01405862

Tilson, J. K., Sullivan, K. J., Cen, S. Y., Rose, D. K., Koradia, C. H., Azen, S. P., & Duncan, P. W. (2010). Meaningful gait speed improvement during the first 60 days poststroke: Minimal clinically important difference. *Physical Therapy*, *90*(2), 196–208.

Tjaden, K., Sussman, J. E., & Wilding, G. E. (2014). Impact of clear, loud, and slow speech on scaled intelligibility and speech severity in Parkinson's disease and multiple sclerosis. *Journal of Speech, Language, and Hearing Research*, *57*(3), 779–792. https://doi.org/10.1044/2014_JSLHR-S-12-0372

Turner, D., Schünemann, H. J., Griffith, L. E., Beaton, D. E., Griffiths, A. M., Critch, J. N., & Guyatt, G. H. (2010). The minimal detectable change cannot reliably replace the minimal important difference. *Journal of Clinical Epidemiology*, *63*(1), 28–36. https://doi.org/10.1016/j.jclinepi.2009.01.024

Turner, M. R., Brockington, A., Scaber, J., Hollinger, H., Marsden, R., Shaw, P. J., & Talbot, K. (2010). Pattern of spread and prognosis in lower limb-onset ALS. *Amyotrophic Lateral Sclerosis*, *11*(4), 369–373. https://doi.org/10.3109/17482960903420140

van Brenk, F., Stipancic, K. L., Kain, A., & Tjaden, K. (2022). Intelligibility across a reading passage: The effect of dysarthria and cued speaking styles. *American Journal of Speech-Language Pathology*, *31*(1), 390–408. https://doi.org/10.1044/2021_AJSLP-21-00151

Van Nuffelen, G., De Bodt, M., Vanderwegen, J., Van De Heyning, P., & Wuyts, F. (2010). Effect of rate control on speech production and intelligibility in dysarthria. *Folia Phoniatrica et Logopaedica*, *62*, 110–119. https://doi.org/10.1159/000287209

Weir-Mayta, P., Spencer, K. A., Eadie, T. L., Yorkston, K. M., Savaglio, S., & Woollcott, C. (2017). Internally versus externally cued speech in parkinson's disease and cerebellar disease. *American Journal of Speech-Language Pathology*, *26*(2S), 583–595.

Yorkston, K. M., Beukelman, D., Hakel, M., & Dorsey, M. (2007). *Speech Intelligibility Test (SIT) for Windows [computer software]* [Computer software]. Institute for Rehabilitation Science and Engineering at Madonna Rehabilitation Hospital.

Yorkston, K. M., & Beukelman, D. R. (1981). Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate. *Journal of Speech and Hearing Disorders*, *46*(3), 296–301. https://doi.org/10.1044/jshd.4603.296

Young, B. A., Walker, M. J., Strunce, J. B., Boyles, R. E., Whitman, J. M., & Childs, J. D. (2009). Responsiveness of the Neck Disability Index in patients with mechanical neck disorders. *The Spine Journal*, *9*(10), 802–808. https://doi.org/10.1016/j.spinee.2009.06.002

Young, V. N., Jeong, K., Rothenberger, S. D., Gillespie, A. I., Smith, L. J., Gartner-Schmidt, J. L., & Rosen, C. A. (2018). Minimal clinically important difference of voice handicap index-10 in vocal fold paralysis. *The Laryngoscope*, *128*(6), 1419–1424. https://doi.org/10.1002/lary.27001