MURAJ
Minnesota Undergraduate Research & Academic Journal

z.umn.edu/MURAJ

Comparing two methods for measuring speech intelligibility in two different environments

Jacob Ender

Abstract

Purpose: Speech intelligibility tests evaluate the proportion of a speech signal produced by a speaker that is understood by a listener. The touchstone for measuring speech intelligibility is orthographic transcription (TRA). This study evaluated the validity, reliability and efficiency of a subjective measure, visual-analog scaling (VAS) compared to TRA in two listening environments, in-lab setting (ILS) and outside-lab setting (OLS). This was accomplished using a small group of speakers with increasing levels of speech severity.

Methods: Sixty-five participants listened to recordings of 13 unique sentences from each of the six speakers. Each participant transcribed the sentence of 3 speakers and rated the sentences of 3 different speakers with VAS. Approximately half the participants were in the ILS environment and the other half in the OLS environment. Stimulus presentation and recording of responses was done with a specially designed browser-based application.

Results: Good agreement was observed in the aggregated TRA condition in the ILS and OLS environments. Although there was reasonable agreement between the aggregated TRA and VAS scores in both environments, there was much greater variability between participants using VAS. Administration of VAS was nearly 4 times faster than TRA.

Conclusion: Analysis of this small group of speakers suggests further study is required to determine if VAS may be a useful clinical tool in lieu of transcription. Internet-based transcription of disordered speech distributed across clinicians remains plausible.

Introduction

Speech intelligibility tests are commonly used in clinical settings to assess individuals affected with speech disorders, such as dysarthria. Speech intelligibility refers to the proportion of the speech signal transmitted by the speaker that is understood by a listener (0% represents no comprehension, 100% represents complete comprehension). Currently, TRA is used with the only standardized assessment for measuring speech intelligibility in people with dysarthria, which is the Sentence Intelligibility Test.1 While speech intelligibility tests have proven to be useful in determining speech impairment severity levels, functional limitations, and monitoring change in individuals over time, some of the methods of assessing speech intelligibility have shown to be unreliable. There are two general methods for assessing intelligibility: transcription and rating. Speech intelligibility rating is clinically significant because it allows clinicians to determine the severity of impairment in speakers with communication disorders and provides a basis for monitoring change during and after the course of therapy.[2]

Transcription entails a blind (naive) listener writing down word for word what a person says. This method requires that one clinician selects the stimuli (words or sentences), presents the stimuli and audio records the responses from a participating speaker. A second clinician, blinded to the participating speaker, must transcribe and score the recording. This method is considered the most valid and most reliable method of assessing speech intelligibility, but the demand on time and resources are substantial.[1] Subjective rating scales such as equal-appearing-interval scales or estimating the percentage of intelligible speech have been used in the past. Although these subjective rating scales require only one clinician to administer and are 4 times faster, they have substantially reduced intrajudge and interjudge reliability.[1]

VAS has recently been studied as a subjective means to rate intelligibility and showed reasonable agreement with transcription.[3,4] For VAS, listeners rate a speaker's intelligibility by using a continuous scale to estimate how much of the content they were able to understand. Because of the subjective nature of VAS - and the limited amount of studies conducted to determine the reliability of VAS, there is still much to be understood.

However, in an earlier investigation by Tjaden et al, it was found that the source of increased intelligibility in speakers with Parkinson's disease under different verbal directives reflected strong agreements between VAS and TRA, but less-so in the mild to severe range of intelligibility ratings.[5] The findings of Tjaden et al show promise for the potential use of VAS for assessing intelligibility and reflect the need to test the reliability and validity of VAS when compared to TRA in a range of different conditions. In this sense, our study expands on Tjaden et al by testing VAS in the two conditions: In-Lab Setting and Outside-Lab Setting.

Ideally, having more than one rater per client per session would help to yield the most valid results.[1] Because of the need to guard against familiarity of the raters to the speaker, having multiple raters would improve the validity of scoring by assuring regression to the mean. In clinical settings, having multiple clinicians engaged in a single intelligibility task may simply not be possible. Exploring the possibilities of internet browser-based applications that could help remote clinicians utilize TRA or VAS may alleviate time and resource constraints and assure a more valid measure by making a wide breadth of available professionals available online.

This investigation entailed a retrospective analysis of previously collected data and aimed to answer the following questions.

(1) Is there a difference between TRA and VAS when measuring speech intelligibility in persons with a wide range of speech severity?

(2) Is there a difference in scoring using TRA and VAS between the ILS and OLS environments?

(3) What is the difference in time taken to score TRA and VAS in the ILS and OLS environments?

Method

Participants

Sixty-five volunteer listeners participated in TRA scoring and VAS rating of audio files of spoken sentences, 32 for the in-lab setting (ILS), and 33 for the outside-lab setting (OLS). The participants ranged in age from 18 to 35 years, and they all reported having normal hearing and were without neurological or communication impairments. They were all directly consented under the supervision of the University of Minnesota's IRB. No attempt was made to balance for gender in or between the two groups.

Stimuli

The audio stimuli used for measuring intelligibility were selected from the TORGO database.[6] This database contains articulatory-kinematic and acoustic data from eight adult speakers with dysarthria and seven adult speakers without dysarthria. Six speakers with dysarthria, ranging from mild to severe impairment, were selected from this database; SIT audio recordings from two of the speakers with dysarthria were judged to be of poor quality and were not included in this experiment. The audio data were sampled at 16 kHz. Each speaker read aloud a number of items, including 162 sentences from the Sentence Intelligibility Test (SIT), and eleven unique sentences for each speaker were selected.[3] The eleven sentences ranged from 5 to 15 words in length. One common sentence, not from the eleven unique sentences, was added to each speaker's set of recorded sentences. This was considered as a means to evaluate *interjudge* reliability. In addition to the eleven unique sentences and one common sentence, one last additional sentence was included during the presentation of each speaker's eleven sentences. This sentence was selected prior to each data collection session and was selected randomly from each speaker's set of eleven to be repeated. The repetition of this sentence was always at least one sentence removed from the first presentation. The repeated sentences within each speaker's audio recordings were intended to be used to analyze *intrajudge* reliability. Thirteen sentences for each speaker were presented to each participant for TRA or VAS. This resulted in 78 total sentences that were evaluated.

Procedure

The software to present the stimuli and record responses for this experiment was developed by LATIS (Liberal Arts Technologies and Innovation Services) at the University of Minnesota - Twin Cities, in consultation with the principal investigator. The experiment was designed to be distributed over the internet via browser, specifically by Chrome™ or Firefox™.

When participants logged on to the experiment, they were first asked to read a consent form and indicate by checking a box if they wished to participate. If they answered no, the session was immediately terminated; if they answered yes, the experiment began. Before data presentation was initiated, the software determined which three speakers would have their speech transcribed and which three would have their speech rated with VAS. The order of TRA or VAS was determined randomly for the first participant and then counterbalanced for the remaining participants. Both methods scored each of the three speakers one at a time.

Before a participant began one of the two methods (TRA or VAS), a brief practice period was given before each procedure. The practice period consisted of presenting three different sentences from three different speakers with dysarthria (mild, moderate, severe) and asking the participant to transcribe or rate the sentences using VAS. The speakers and sentences were not from the TORGO database but were instead sourced from the Nemours database.[7] After scoring each sentence, the participant was shown what the speaker said, so they could judge the accuracy of their response.

The instructions for listening to each sentence conformed with directives provided in the manual for the Standardized Intelligibility Test reference. The first time a sentence was presented it played through completely - all controls and ability to operate the browser were locked during this period. The participant could then respond either by typing out what they heard (TRA) or rate using a scale (VAS), depending upon the pre-selected condition. After the initial presentation of the stimuli, the listener had the option to continue to the next sentence or hear the sentence played again. If they chose to repeat a sentence, they could pause playback of the recording at any point and then proceed until it had finished playing. After this, all control for play and pause were locked and participants had to proceed to the next sentence.

There were two relative locations for the experiment: in-lab setting (ILS); and outside-lab setting (OLS). The ILS participants were seated in a sound-treated booth, wearing high-fidelity headphones (Sennheiser, HD 280). Digital audio files were delivered to the headphones at 44.1 kHz by USBpre DAC (audioPro). The OLS participants were instructed to do the task in a quiet environment, use any digital device that would run the browser application, and to wear headphones or in-the-ear buds.

All participants' responses were stored on a secure server. The data recorded included the typed transcription and VAS ratings for intelligibility, the time in seconds to respond to each sentence, if the participant replayed a sentence again, how many times the participant paused the sentence the second time it was played, and which sentence in a set was repeated and where it occurred in the set.

For the retrospective interpretation, each listener's scoring of speakers in the two conditions (TRA and VAS) and two sources (ILS and OLS) were aggregated into averages for each speaker. Scores for each speaker were used to compare performance in each combination of sources and conditions among listeners. Overall time differences between conditions and sources among all listeners were also compared. Data visualization and analysis included box plots displaying average scored intelligibility (%) and average time to score by condition (TRA and VAS) and environment (ILS or OLS). Two-way ANOVA analysis was performed on both the percentage of intelligibility scores and timing data of condition and environment. Regression analysis was performed to examine the relationship between VAS by TRA and ILS by OLS. Intraclass correlation was performed to access the agreement between participants by condition and environment.

Results

Grand Mean Values for Intelligibility

Table 1 shows the grand means for percent intelligibility for the 6 speakers collapsed for listeners, conditions, and environments. Speakers are ranked in order from most intelligible (Speaker 1) to least intelligible (Speaker 6). Speakers 1 and 2 scored near 100% and have been

	Speaker 1	Speaker 2	Speaker 3	Speaker 4	Speaker 5	Speaker 6
	98.91	97.46	75.82	63.20	54.65	28.87

Table 1: *Intelligibility (%) aggregated for TRA and VAS scoring conditions in the ILS and OLS Environments for the Six Speakers with Dysarthria.*

excluded from further presentation of results and discussion of their data for scoring intelligibility.

Correlation between Transcription and VAS

Figure 1 shows two bivariate regressions created to examine the correlation between scores for transcription and VAS, for both ILS (left plot) and OLS (right plot). The grand means for TRA and VAS are shown for the 4 speakers in both plots. The R2 values for both settings are higher than .95, indicating a strong correlation, and demonstrating that mean transcription scores are strong predictors of VAS scores. Between ILS and OLS, the difference between R2 values was observed to be less than .01, which indicates that setting has little effect on scoring variability.

Analysis of Intelligibility by Conditions and Environments

Figure 2 displays box plots for Speakers 3, 4, 5, and 6. The box plots show the listeners' average percent intelligibility for TRA and VAS conditions for the ILS and OLS environments. Within each plot the two boxes to the left are for ILS and the two to the right are for OLS. In each pair the brown color represents the TRA condition and the blue color represents the VAS condition. The horizontal line in each box represents the second quartile, while the lower and upper limits of the box show the first and third quartiles, respectively.

A three-way ANOVA of Speaker by Environment by Condition revealed two-statistically significant main effects and two interactions (see Table 2). The two significant main effects were for Speaker and Condition and the two significant interactions were Speaker x Condition and Speaker x Environment x Condition. A Tukey HSD post-hoc analysis showed that all pairings of speakers were significantly different from each other (p < .01). This was expected because the four

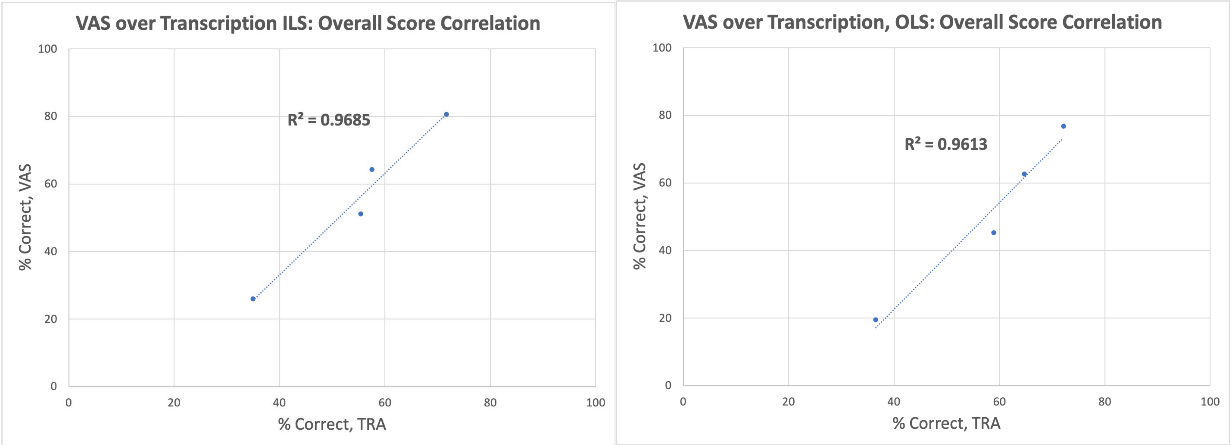

Fig 1: *Correlation of scores between TRA and VAS for ILS and OLS.*

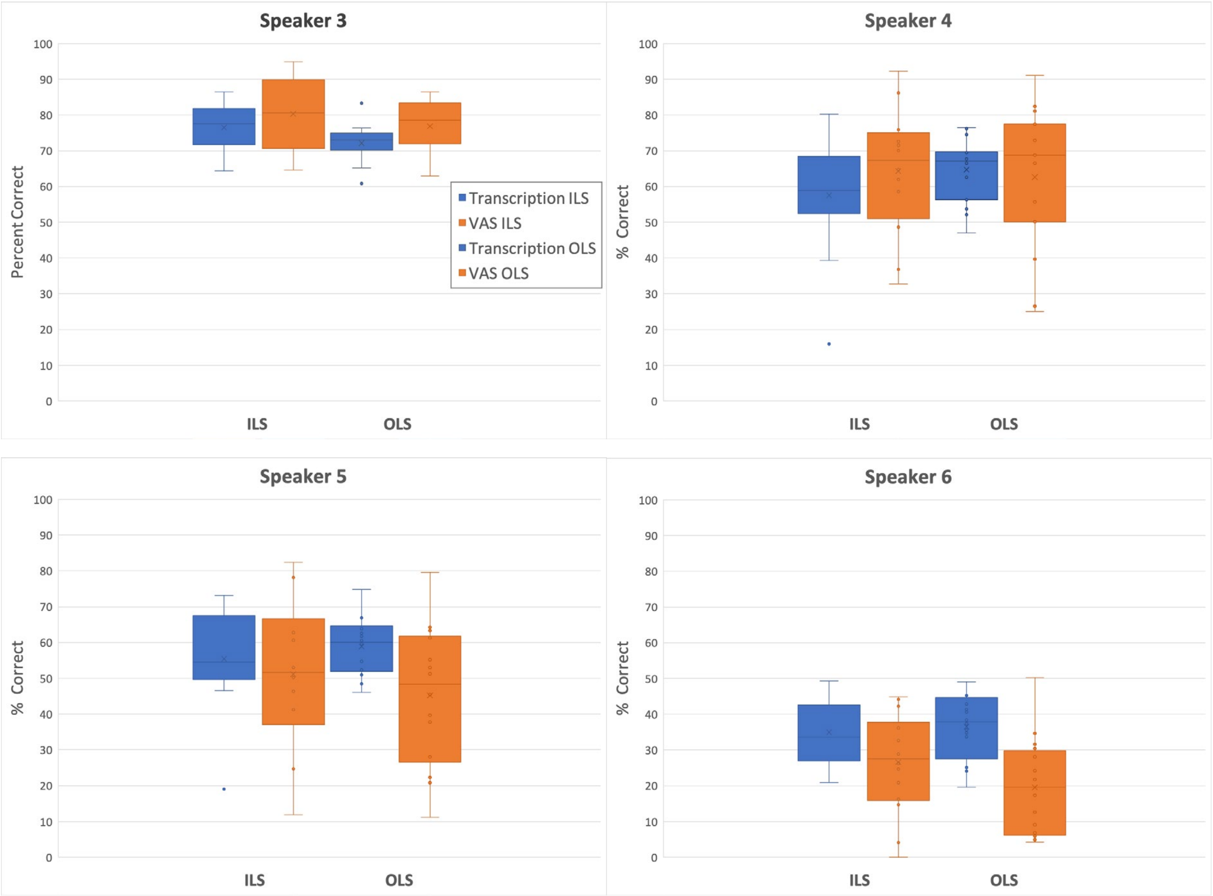

***Fig 2**: Boxplots of scores for transcription and VAS in the ILS and OLS settings for speakers 3, 4, 5, and 6.*

speakers with dysarthria were chosen because of the range of speaking impairment. Because of the three-way interaction it was decided that a two-way ANOVA (condition by environment) would be performed individually for Speakers 3, 4, 5, and 6. Before breaking down each speaker's data, some general observations of the data can be seen. First, there is reasonably good overlap of the TRA conditions in the ILS and OLS environments for each speaker. Second, with the exception of Speaker 6, there is some overlap between the TRA and VAS scoring, but with a much greater range of average scores for the VAS condition. For Speakers 3, 4, and 5 there were no statistically significant differences between TRA and VAS, environments ILS and OLS, and with no interactions. For Speaker 6, the ANOVA revealed only a statistically significant difference between conditions of TRA and VAS [F (1,59) = 21.58, p < .001, $\eta p^2 = 0.26$].

Time to Score

Figure 3 shows boxplots for the four selected speakers, showing the average and distribution of time (in seconds) per item to score for the TRA and VAS conditions, between the ILS and OLS environments. The data shows that VAS took the least amount of time in every instance. A trend in the data revealed that both TRA and VAS took less time in the OLS environment than in the ILS. Speakers 1 and 2 had scores of intelligibility close to 100% and took much less time overall to transcribe and score compared to other speakers. These observations were consistent to the three-way ANOVA Speaker x Environment x Condition (see Table 4), showing three main effects for

	Df	F value	Pr (>F)	η_p^2
Speaker	3	129.962	< 2.00E-16	0.62
Environment	1	0.229	0.633	0.001
Condition	1	4.928	0.0274	0.02
Speaker:Environment	3	0.197	0.8982	0.002
Speaker:Condition	3	7.762	5.74E-05	0.089
Environment:Condition	1	4.415	0.0367	0.018
Speaker:Environment:Condition	3	0.031	0.9926	0
Residuals	239			

Table 2: *Three-way ANOVA for Speakers 3, 4, 5, 6 by Environment and Condition*

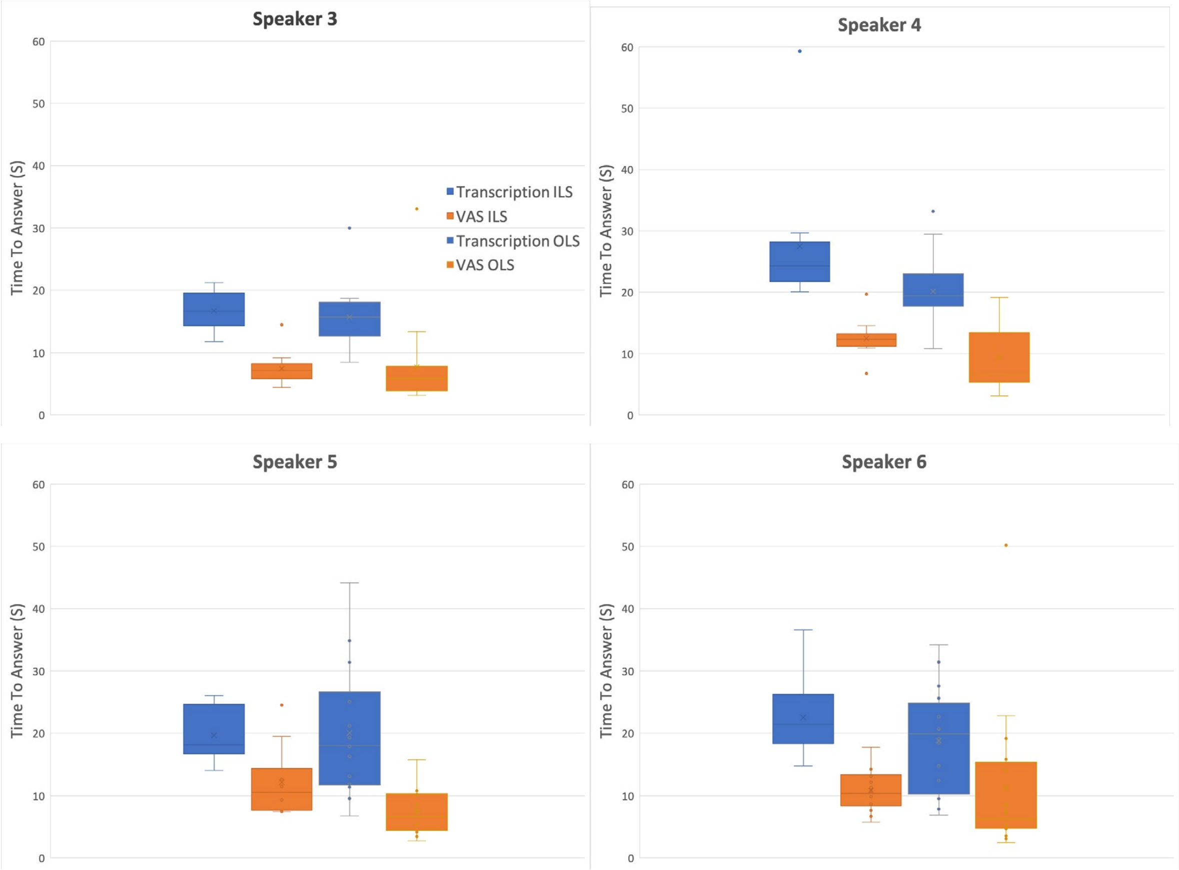

Figure 3: *Time to answer in seconds for each using transcription and VAS for speakers 3-6.*

	Speaker 1	Speaker 2	Speaker 3	Speaker 4	Speaker 5	Speaker 6
	9.97 (6.24)	8.00 (5.75)	11.80 (6.35)	15.20 (7.71)	16.10 (9.19)	15.80 (9.37)

Table 3*: Average time and standard deviations, in seconds, aggregated for TRA and VAS scoring conditions in the ILS and OLS environments for the Six Speakers with dysarthria.*

	Df	F Value	Pr(>F)	η_p^2
Speaker	5	23.12	<2.00E-16	0.244
Environment	1	8.04	0.00484	0.022
Condition	1	295.583	<2.00E-16	0.452
Speaker:Environment	5	0.795	0.55393	0.011
Speaker:Condition	5	1.663	0.14281	0.023
Environment:Condition	1	1.616	0.2045	0.004
Speaker:Environment:Condition	5	1.734	0.12598	0.024
Residuals	358			

Table 4: *Three-way ANOVA for average time to score each item for the 6 speakers.*

Speaker, Environment, and Condition and no interactions.

The average time to transcribe, aggregated for environment and speakers was 17.7 seconds per item. This value times 13 items equals approximately 230 seconds or 3.9 minutes per speaker. Added to this was the time it took to score each item against the key to determine the number of correct words heard and on average took 3 minutes. It took a total of 6.9 minutes to determine the % intelligibility per speaker by transcription. The overall average time per item for the VAS, which did not require further scoring, was 7.8 seconds and when multiplied by 13 items equals 101.4 seconds or 1.69 minutes. On average, VAS was four times faster than transcription.

Discussion

This study sought to determine if rating using visual analog scaling (VAS) was equivalent to transcription in measuring intelligibility in two conditions, In-Lab Setting and Outside-Lab Setting. The speech severity of these speakers ranged from mild to severe. When considering the three research questions, the ILS TRA data was used as a benchmark for comparison.

Research Question 1: Is there a difference between transcription and VAS when examining speech intelligibility in persons with a wide range of speech severity?

TRA vs. VAS ILS: Mean scoring performance between TRA and VAS in the ILS varied by as little as 2% (as in speaker 3) and as much as 9% in speaker 4. This observation is in accordance with the excellent agreement shown in the regression analysis.

Although average intelligibility measures showed reasonable agreement between VAS and TRA, the average scoring distribution, as shown by the interquartile range (IQR), was greater for VAS. The average interquartile range was nearly 1.25 times greater for the VAS in the ILS environment, and was nearly double for VAS compared to TRA for the more affected speech of Speakers 5 and 6.

TRA vs. VAS OLS: Transcription and VAS had the largest differences in mean and distribution spread in the OLS. The means of transcription and VAS differed by as much as 19%, and as little as 2%. In terms of distribution spread of the interquartile range, there is much less overlap between the conditions in the OLS than the ILS. In speaker 6, there is almost no overlap of agreement between transcription and VAS in the OLS.

For every speaker in both settings, the coefficient of determination with VAS as a function of transcription was calculated to be $R^2 = .87$ or greater. This coefficient shows the scores for transcription are good predictors of VAS scores. Between the ILS and OLS for each speaker, R^2 values differed by an average of .0825; the lowest difference occurring between the settings for speaker 3, and the largest difference between the settings for speaker 6.

Research Question 2: Is there a difference in scoring using transcription and VAS between an in-lab condition (ILS) and an outside-lab condition (OLS)?

In Figure 1, the bivariate regressions of correlation between scores using Transcription and VAS show that there is little difference in correlation between ILS and OLS; less than .01 difference was observed between R^2 values. Stipancic et al found that subjective intelligibility scores in the form of percent estimates were lower than scores derived from a transcription task for speakers with dysarthria.[3] In our study, it was also observed that listeners tended to underestimate speaker intelligibility, especially when intelligibility was 60% or lower. In regard to varying levels of intelligibility, Mocarski et al found that in testing speech intelligibility in background noise with varying levels of signal-to-noise ratios, VAS scoring had higher levels of variability when speakers were in the mid-severity range, and that variability drastically decreased as a speaker's intelligibility level approached either completely unintelligible or completely intelligible.[2]

Transcription ILS vs. OLS: Transcription in the ILS source best represents the standard of clinical practice in measuring speech intelligibility. Overall, transcription in the ILS and OLS yielded a similar distribution within the interquartile range. In every speaker besides speaker 3, however, transcription in the OLS yielded higher mean scores. Speaker 3 is the only speaker where listeners using transcription in both the ILS and OLS did not have a similar mean and distribution.

VAS ILS vs. OLS: In every speaker, means of VAS scoring in the ILS vs. OLS vary by 9% or less. Distribution sizes within the interquartile range are very similar overall. However, OLS scores for VAS seem to consistently underestimate the speaker's performance, especially in speakers where a relatively lower level of speech intelligibility across sources and conditions was confirmed. Listeners also overestimate speaker intelligibility where listeners across sources and conditions agreed that the speaker had a relatively higher level of intelligibility.

The wide range of distribution in the interquartile range of TRA in the ILS and its mean should be considered a comparison measure to the other sources and conditions. When comparing the TRA ILS mean to the means across sources, conditions, and speaker, the means differ by 9% on average. Differences are found as low as 7% as observed in speakers 3 and 5, and as much as 12% as observed in speaker 6. In another study that researched transcription and VAS use in sentence intelligibility tests with variable listener exposure, Abdur et al determined that anything below a 7% difference was within a range that could be deemed allowable variation, and anything above 7% denotes a clinical change, which may or may not be accurate.[4] In examining the total range of scoring for each speaker across all combinations of sources and conditions, the smallest range was found in speaker 4, with a range of 14%. The largest range was found in speaker 5, with a range of nearly 40%.

Research Question 3:What is the difference in time taken to answer using transcription and VAS between an inside- lab condition (ILS) and an outside-lab condition (OLS)?

In both conditions, using VAS took significantly less time when compared to transcription by an average of nearly 10 seconds. Among all participants, the interquartile range of transcription and VAS when comparing the rating systems as they were used in the two conditions had comparable distributions in terms of range and interquartile size.

TRA: ILS vs. OLS: Transcription done in the ILS and OLS were very similar when looking at the time taken to answer. In every instance, the mean time for OLS fell slightly lower than the ILS. For ILS, the interquartile range tended to stay within a range of 7 seconds. For OLS, the interquartile range varied widely, such as in the cases of speaker 3 and 5 - which were the speakers rated as the least intelligible of the six.

VAS: ILS vs. OLS: As a trend observed throughout all speakers, VAS performed in the OLS typically took less time to give a rating in every instance. The interquartile range for VAS rating in both ILS and OLS took significantly less time to perform.

TRA vs. VAS in ILS: With a mean time near 18 seconds to perform transcription in the ILS, VAS outperformed transcription by just over 10 seconds. The distribution in transcription is much more consistent than VAS which had several outliers.

TRA vs. VAS in OLS: Transcription and VAS were distributed less consistently in the OLS. Both sources had significant amounts of outliers, with larger ranges in the fourth quartiles than in the first quartile.

Conclusion

Through this study it has been shown that TRA produces similar results in the ILS and OLS settings. It was also found that statistically, transcription and VAS are similar . Our study expanded on this by demonstrating that those similarities are persistent in both ILS and OLS. In light of this, browser-based applications to rate speech intelligibility become more feasible. In the future when more studies expand on this finding - and if results support the use of VAS in different settings - using browser-based applications have the capability of creating an online source of professional resources to which time and resource-strained clinics could turn to for the important clinical diagnostic use of intelligibility testing.

One main difference between the rating systems is that VAS demonstrated more variability in scoring in both the ILS and OLS. Approaches to future research studying VAS as a possible clinical tool should make an attempt to address reasons for this variance. One confounder to be considered in new studies is listener motivation. Future studies could recruit experienced clinicians to perform the tasks or create techniques to further motivate listeners to examine if motivation has a significant effect on the amount of observed variance. In other words, what happens when you are able to motivate individuals to score more accurately?

Our study also utilized stimuli from only six speakers with only one specific form of dysarthria - spastic dysarthria secondary to cerebral palsy. Future studies should expand on how VAS rating compares to transcription when stimuli from a greater number of speakers with various types of dysarthria are incorporated. Different types of dysarthria produce a range of speech deficits including aberrant voice quality and articulatory impairments that may affect the agreement and reliability between TRA and VAS in ways not observed in this study.

References

1. Yorkston, K. M., & Beukelman, D. R. (1978). A comparison of techniques for measuring intelligibility of dysarthric speech. *Journal of communication disorders*, *11*(6), 499-512.

2. Mocarski, S. T., & Watson, P. J. (2016). Reliability and time to administer visual analog scaling of intelligibility. *The Journal of the Acoustical Society of America*, *139*(4), 2048-2048.

3. 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.

4. Abur, Defne, Enos, Nicole M., and Stepp, Cara 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.* 58, 1222-1232.

5. Tjaden, Kris, Kain, Alexander, & Lam, Jennifer (2014). Hybridizing Conversational and Clear Speech to Investigate the Source of Increased Intelligibility in Speakers With Parkinson’s Disease. *Speech, Language, and Hearing Research*, *57*(4), 1191-1205.

6. Rudzicz, F., Namasivayam, A. K., & Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, *46*(4), 523-541.

7. Menendez-Pidal, X., Polikoff, J. B., Peters, S. M., Leonzio, J. E., & Bunnell, H. T. (1996, October). The Nemours database of dysarthric speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96* (Vol. 3, pp. 1962-1965). IEEE.