**Proceedings of the Institute of Acoustics**

# WORD SCORE VS. STI TESTS OF A PUBLIC ADDRESS SYSTEM IN AN UNDERGROUND TRAIN STATION PLATFORM

O Fong Acoustic Directions, Australia
C Hanrahan-Tan Acoustic Directions, Australia
G Leembruggen Acoustic Directions, Australia

## 1 INTRODUCTION

A mismatch was found by the authors between measured speech transmission index (STI) results using STIPA measurements and the STI derived from measured 1000-word phonetically-balanced scores on the platform areas of a new underground train station in Sydney. This paper presents the details and results of the testing that was undertaken on the platform and provides a discussion of the reasons that may have led to the difference in results between the measured STIPA values and those calculated from PB-word score relationships.

The project specification required the intelligibility on the platforms to meet a minimum STI of 0.5 or Common Intelligibility Scale (CIS) rating of 0.7 as per Australian Standard AS 1670-4:2018[1]. This level of intelligibility is required in areas that have a signal-to-noise ratio of 10 dB or more between the broadcast speech and the background ambient noise, and a reverberation time of less than 1.5 seconds.

During the project's commissioning phase, initial STIPA benchmarking on the platform area showed non-compliance with the project's STI requirements and we were asked to assist with improvements. As the PA system was already installed, the system could only be optimised using equalisation and broadband level adjustments. Equalisation was undertaken to improve the perceived intelligibility and the STI in the presence of ambient noise. Although substantial equalisation, particularly at high frequencies can be used to maximise STI, the resulting tonal balance can actually degrade perceived intelligibility. As it was important to maintain natural voice tonality for listening comfort and perceived intelligibility, equalisation was undertaken to achieve these outcomes.

Measurements made on the platforms after optimisation showed that the average STI was still unlikely to meet the project's intelligibility requirements over the entire platform. This was due to a number of factors including i) the type of speaker used and their placement, ii) the restricted system architecture and rudimentary nature of the digital signal processing (DSP) and iii) architectural acoustics including limited sound absorption in the lateral direction and large parallel reflective surfaces. As conformance to the project's specification was critical, the client suggested to attempt word score testing as an alternative method to demonstrate compliance.

The phonetically-balanced-word (PB-word) score method using 1000 words was used to test for the percentage of words that was correctly heard. This percentage of correct words was then converted to its equivalent CIS and STI rating according to the conversion graph provided in AS 1670-4:2018, which is based on the original paper by Barnett & Knight[2]. The CIS and STI results obtained through PB-word scores complied with the project requirements.

Letowski and Scharine[3] provides a comprehensive discussion of various word and sentence tests and conclude that the phonetically balanced words are the most accurate of all the standardized tests and is recommended for use when high data accuracy and sensitivity are required.

## 2 TEST METHOD

### 2.1 Word Score Testing

Word score testing was undertaken on two separate days in June and July 2024. As the station staff were busy preparing for handover and test trains were running, the listening test had to work around these intrusions.

The test process broadly adopted the requirements provided in standard ANSI/ASA S3.2-2020[4] relating to the setup and administration of a word score test. In particular, this included the acoustic conditions at talker location, acoustic conditions at listener location, spoken test materials, and the selection and training of talkers to record the lists as per Section 6 and 7 of the ANSI/ASA standard. Using professional recording equipment, an experienced male voice-artist with an Australian accent recorded twenty lists of fifty phonetically-balanced words in a suitable recording studio. A carrier phrase ("Please write the word…") preceded each test word and a seven second gap was provided between sentences. This gap allowed test subjects to write down the word and minimise cross-talk between sentences in the platform's acoustic environment.

The use of a single voice-artist as talker is a deviation from the recommendations in ANSI/ASA S3.2-2020, which recommends a minimum of five talkers. The use of only one talker would likely reduce the repeatability of the results. However, as the testing was undertaken as part of a commercial project, the engagement of an additional four voice artists would be costly and there was no requirement to repeat the test.

Forty-seven test positions were marked out along the ground to cover the entire platform area, These test positions remained identical across both test days. At each location, the word score was calculated by counting the total correctly-identified words over the two days and dividing them by the total number of presented words.

The measured spatially-averaged ambient noise level on the platform was 61 dB ($L_{Aeq}$) and the broadcast speech level was approximately 76 dB to 80 dB ($L_{Aeq}$) during testing. These levels were measured as a spatial average of approximately one quarter of the platform area. The pre-recorded word lists were preloaded to the station's PA system and broadcast from the Station Masters Room to the loudspeakers serving the platform area. There was no contractual requirement to assess the STI with the tunnel ventilation system (TVS) operating, however word score tests conducted at ten positions with the TVS operating showed little change from the no TVS state.

The selection of test subjects generally followed the recommendations in ANSI/ASA S3.2-2020. Six listeners were found to participate in the word score testing. All of them were native English speakers and self-reported to have normal hearing. Five of these listeners were present for both test dates; however, one listener changed between test dates. Each listener undertook the test in 20 positions, which equated to 1000 words per listener.

The test process on both test days was as follows:

a) The six participants each took a position along the platform.

b) One of the set lists was broadcast through the platform's PA system while participants recorded their answers with pen and paper on a clipboard.

c) After the reading of each list was complete, the response sheets were handed over to markers who counted the percentage of correct words. The results were cross-checked between markers for instances of uncertainty and error.

d) The results were compiled and recorded after each marking session.

e) Participants moved to six new positions along the platform and the process was repeated until the set number of lists was complete for the day.

f) To avoid listening fatigue, regular short breaks were provided between word lists and a longer break was provided for lunch. Snacks and water were also made available. To manage overall fatigue, four of the participants sat on floor during the testing while the remaining two participants stood.

In one instance there was a short period of construction noise which occurred during one of the tests. This extraneous noise affected two listeners and five of the sentences in one list. To compensate for this, five new replacement words from a different list were used and the affected listeners retested with these words.

## 2.2 STIPA Testing

To compare the measured word scores with STI ratings, STIPA measurements were undertaken in ten representative positions along the platform that had been used for the word score test, These positions were distributed across the platform area and provided a reasonable sample of the platform's STI performance.

A STIPA signal was preloaded to the station's PA system and broadcast from the Station Masters Room to the loudspeakers on the platforms at a sound pressure level of approximately 76 dB to 78 dB ($L_{Aeq}$), which was measured as a spatial average in an area along the platform. Ambient noise levels were low with the TVS not operating and care was taken to ensure measurements were made during periods without trains.

All measurements were made using an NTI XL2 Type 1 acoustic analyser. Calibration checks of the acoustic analyser was made prior to and post measurements to ensure the validity of data acquired by the device. The microphone was at a height of approximately 1.5 m above the ground to represent the listening height of standing people.

# 3 TEST RESULTS

## 3.1 Word Score Test

The average word score of all 47 test positions on the platform was 85%. Using the conversion graph provided in Figure I.1 in AS1670.4:2018, the average word score was converted to the equivalent CIS of 0.75 and STI of 0.57, both of which achieved the project's intelligibility requirement.

Word scores at individual test locations ranged between 73% and 93% correct. These results are shown in Figure 1.

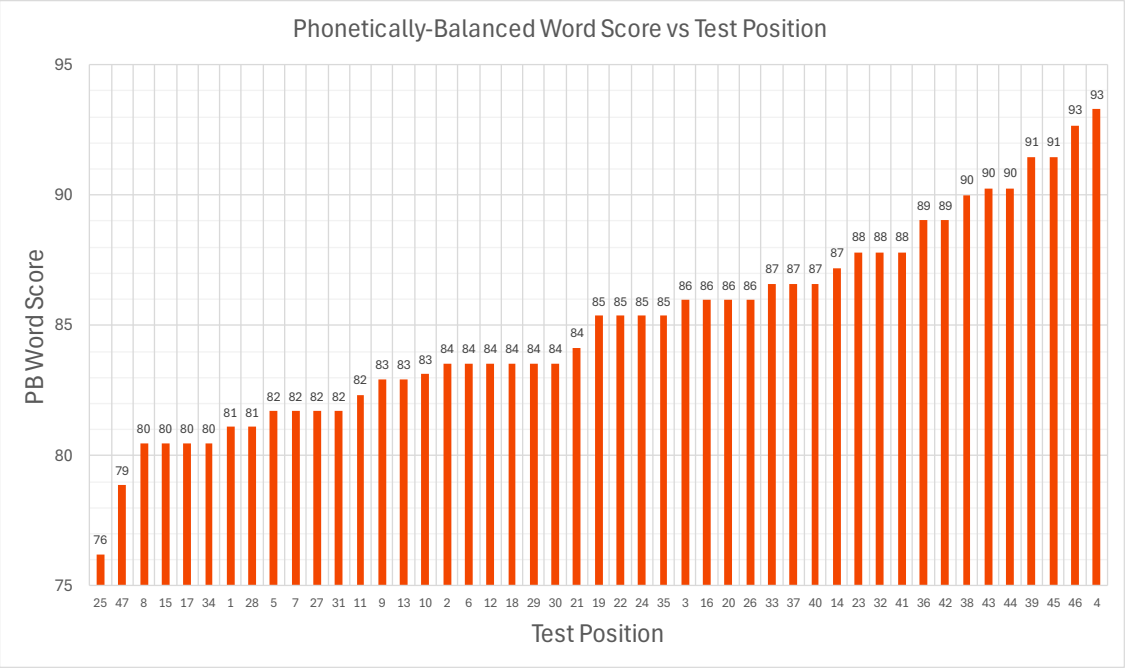


**Figure 1.** Percentage of correct words from word score testing on the platform area.

## 3.2 Comparison of Word Score versus Measured STI

Table 1 presents the pairs of measured and equivalent STIs at the ten locations at which STIPA measurements and word score testing were undertaken. On average, the measured STIs were found to be 0.1 lower than the equivalent STIs derived from the 1000 PB-word score.

Figure 2 presents the individual word scores at each test location as a heat map in combination with the ten STIPA-measured STI ratings and the $L_{Aeq}$ sound level during the STIPA measurement. Note that the entire platform is represented by the grey area and the crosses indicate the test locations.

**Table 1.** Word Score Results vs STI.

| Measurement Position | Measured STI | Calculated STI from PB1000 Word Score | % Correct from PB1000 Word Score |
|---|---|---|---|
| 7 | 0.44 | 0.53 | 82% |
| 12 | 0.45 | 0.57 | 86% |
| 14 | 0.51 | 0.73 | 96% |
| 19 | 0.49 | 0.56 | 84% |
| 23 | 0.47 | 0.60 | 88% |
| 26 | 0.46 | 0.52 | 80% |
| 30 | 0.46 | 0.53 | 81% |
| 37 | 0.44 | 0.52 | 80% |
| 39 | 0.45 | 0.57 | 85% |
| 46 | 0.47 | 0.56 | 84% |
| **AVERAGE** | **0.46** | **0.56** | **84%** |

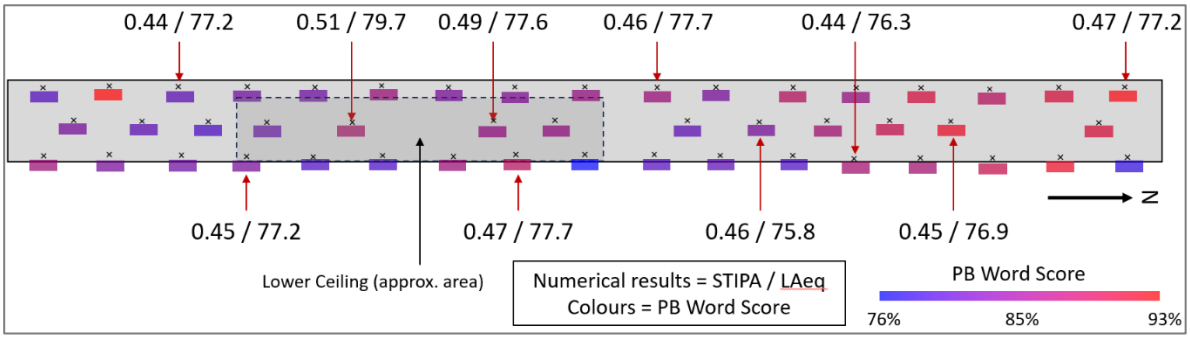


**Figure 2.** A plan showing the location of measurements along the platform area, their associated word scores and, where applicable, the measured STI and sound pressure level measured during STIPA measurement.

# 4 DISCUSSION

## 4.1 General Discussion

A location analysis (see Figure 2) of word score performance across the platform area suggests that i) localised acoustic conditions at test positions and ii) loudspeaker power settings, could be two factors that influenced the word score results. Results in areas with a high and sound-absorptive ceiling near the north end of the platform area provided better word scores than areas with a lowered ceiling. Some test locations also showed a noticeably better or worse word score compared to their directly adjacent test locations, which suggest that, although not obvious from general listening across the platform, there may have been discrepancies between speaker power settings.

An important point to note is that although an overall word score using 1000 PB words was obtained for the combined platform area, the individual word scores at each measured positions were based on a subset of the total words tested. This is due to having insufficient participant numbers for all 47 test points to be occupied during the broadcast of each of the word lists. However, in his discussion

of PB words, Egan[5] notes that the spread of difficulty is approximately the same in each list and each list has nearly the same average difficulty. Although the testing of additional words at all test positions might improve the reliability of the results, the measured word scores still provide a good indication of intelligibility performance at each test position.

Additionally, a streamlined training approach was provided to the test subjects due to project time and cost constraints. The training explained the test process and included a sample of a word list being played. However, no assessment was made of whether the performance of test subjects reached a plateau as recommended in ANSI/ASA S3/2-2020. It is expected that the word score results could possibly improve, especially under difficult listening conditions[5], but that would further increase the difference between the word score and STI results.

## 4.2 Comparison of Results with Literature

As the measured STI using a STIPA signal differed substantially to the equivalent STI from 1000 PB-word scores, an exploration of the relationship between these metrics in AS1670.4:2018 and other alternative relationships in literature was undertaken. The following section compares and discusses each of these relationships with the results obtained on the platforms.

### 4.2.1 Steeneken & Houtgast, 1980

Figure 3 presents the measured results on the platforms with a relationship found in a 1980 study by Steeneken and Houtgast[6]. At all measured locations on the platform, the word scores measured considerably higher than the expected values.

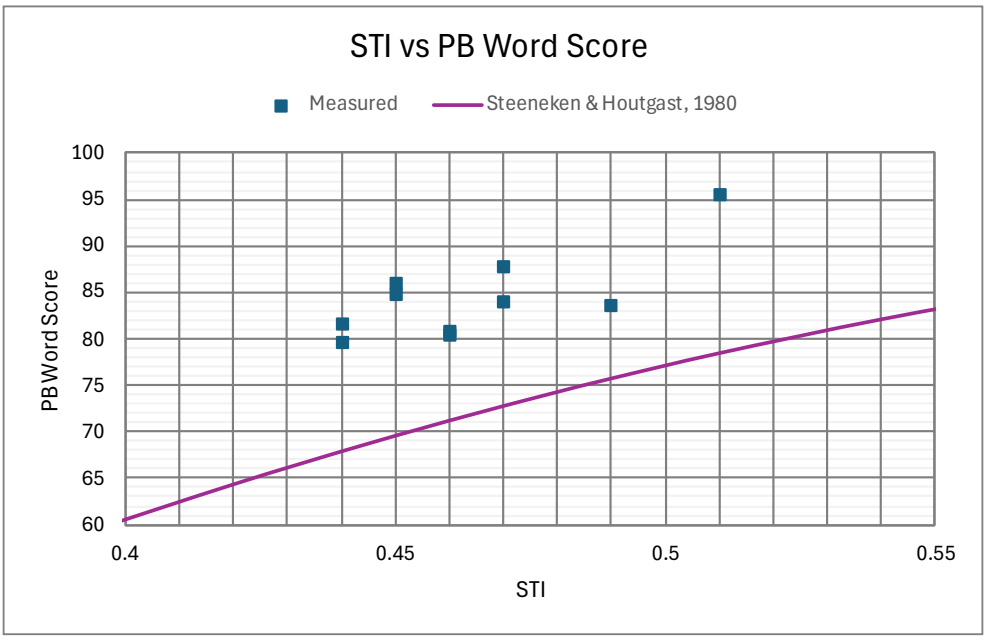


**Figure 3.** A comparison of the measured results on the platform area with the relationship between STI and 1000 PB-word score in Steeneken and Houtgast[6].

Steeneken and Houtgast's relationship between STI and 1000 PB-word score is based on experiments that explored how frequency distortions (noise and bandpass limiting), non-linear distortions (peak clipping), temporal distortion (reverberation and automatic gain control) and digital distortions affect intelligibility.

The relationship between STI and 1000 PB-word score by Steeneken and Houtgast was adopted by Barnett & Knight[2] when developing the CIS rating. This relationship has subsequently been adopted by Australian Standard AS1670-4:2018 and is embedded within the conversion graph provided in the standard between various subjective and objective intelligibility metrics.

One of the differences that has possibly contributed to the mismatch between the test results and the relationship found in Steeneken and Houtgast is the difference in language. Unlike the English words used for the word score testing on the platform area, the Steeneken and Houtgast study uses lists of 50 phonetically-balanced words in Dutch.

Additionally, the Dutch words used by Steeneken and Houtgast are meaningless, whereas the English words used on the platforms have meaning. As nonsense syllables are harder to interpret than words[7], word scores obtained using nonsense syllables would be lower than words with meaning.

The benefits of binaural hearing over monaural hearing could further explain why the word scores are higher in the platform when compared to the expected STI from Steeneken and Houtgast's relationship. Steeneken and Houtgast's experiments were based on monaural speech presented over headsets, whereas the platform tests were undertaken in an acoustic environment that allows binaural listening cues to provide intelligibility benefits. Numerous research has demonstrated the benefits of binaural hearing over monoaural hearing (e.g., Hawley et al.[8], Nábělek & Robinson[9] and Bronkhorst & Plomp[10]).

### 4.2.2 Anderson & Kalb, 1987

In their study to validate the use of STI to model speech intelligibility with the English language, Anderson and Kalb[11] developed an alternative relationship between STI and 1000 PB-word scores. Figure 4 shows this relationship in comparison with the measured results on the platform.

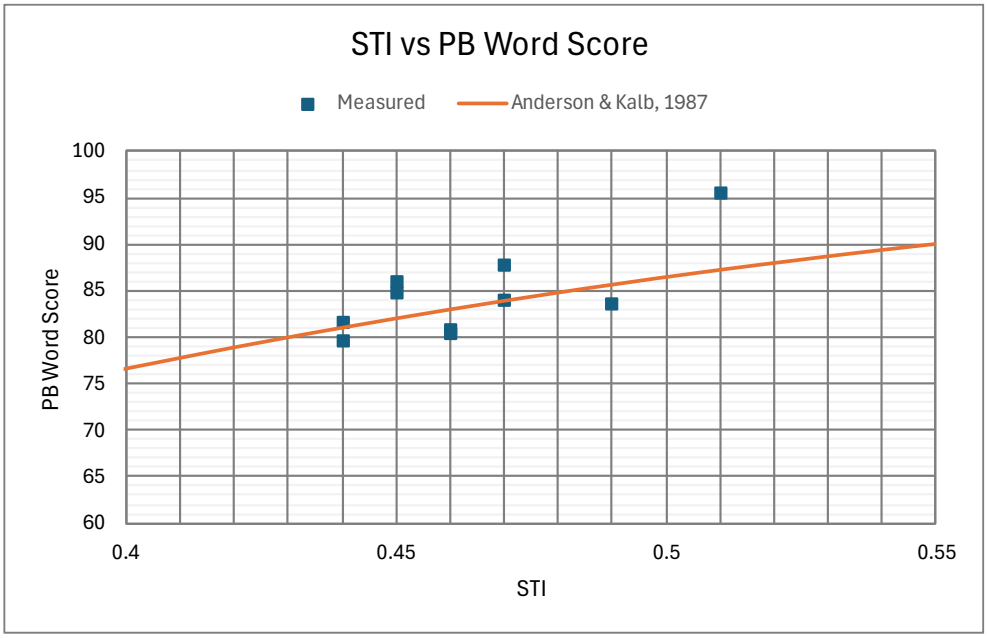


**Figure 4.** A comparison of the measured results on the platform area with the relationship between STI and 1000 PB-word score in Anderson & Kalb[11].

The measured results on the platform fit Anderson and Kalb's relationship better than Steeneken and Houtgast's relationship. A likely contributing factor to this better fit is the fact that both the testing on the platform and by Anderson and Kalb are based on English PB-words.

Similar to Steeneken and Houtgast, Anderson and Kalb explored how noise, band-pass filtering and reverberation affects speech intelligibility in a transmission channel. Anderson and Kalb however did not explore the effects of peak clipping, AGC or digital distortions. Both studies conducted listening tests over headsets and as such did not take into account the benefits provided by binaural cues.

Given the close match found between the platform testing and Anderson and Kalb but not with Steeneken and Houtgast, an argument could be made that the relationship between STI and 1000 PB-word score contained in AS1670.4:2018 should be revised.

### 4.2.3 Morales et al., 2014

Another alternative relationship between STI and PB-word score was found in the study by Morales et al.[12] regarding the verification of STI and intelligibility in reverberant conditions. The study was conducted in a reverberation chamber with five configurations of sound absorbers to produce five reverberation scenarios. A loudspeaker was placed in the room to broadcast PB words for word score testing or to provide an impulse response for indirect STI measurements. A microphone was placed at various locations inside the room with various distances from the loudspeaker to obtain a range of STI values between 0.36 and 0.70. Test subjects were asked to sit on a chair at these test locations, and brief training was provided in which they listened to one of the PB word lists.

Five male talkers were used to generate the broadcast sentences which follow the syntax “Write the word … please”, where the test word is inserted into the ellipsis. Each position was tested by ten subjects listening to only one PB-word list. Assuming that a fifty-word PB list was used, each position had a total of 500 words tested. This is different to the Steeneken & Houtgast and Anderson & Kalb, which both used 1000 words.

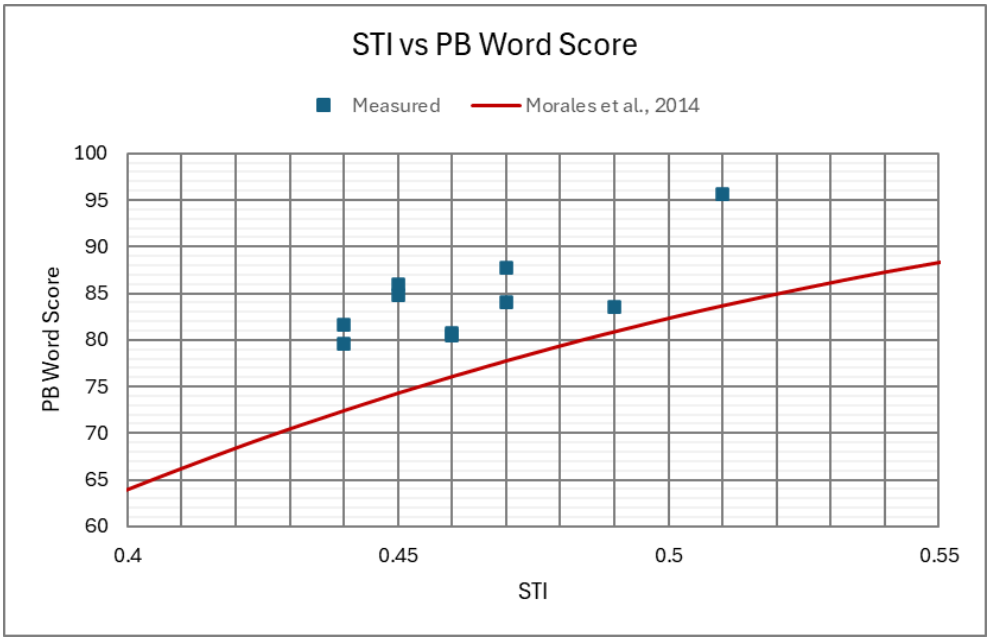


**Figure 5.** A comparison of the measured results on the platform area with the relationship between STI and PB-word score in Morales et al.[12].

As shown in Figure 5, word scores measured on the platform were higher than the relationship from Morales et al. would predict. Compared with the other relationships presented previously, Morales et al. provides a closer match to the test results compared to Steeneken and Houtgast, but the best match is with Anderson and Kalb’s relationship.

It is difficult to conclusively attribute any reason as to why the Morales et al. relationship did not match the results from the platform testing. Morales et al. noted that the difference in training procedure could be a possible reason for their lower word scores compared to the Anderson and Kalb relationship, as test subjects in Anderson and Kalb’s study underwent a more comprehensive training process that ensured they reached a certain level of performance prior to undertaking the test. However, training of test subjects for the word score tests on the platform was streamlined and would have been more similar to the training in Morales et al. than Anderson and Kalb. As such the lack of training in Morales’ work is unlikely the reason for the mismatch.

Another non-reason is the lack of female talkers in Morales et al., as the platform testing also only used a male speech for word delivery to test subjects. As noted by Morales et al., female speech is generally more intelligible than male speech, which accords with the relationship difference between Morales et al. and Anderson and Kalb, which included female talkers. However, Anderson and Kalb stated that they found no significant difference in word scores between speech presented by male and female in their study.

### 4.3 Additional Discussion

It is important to note that as the testing on the platforms was conducted as part of a commercial project, it had to be undertaken under practical time and financial constraints that may not have been present in a laboratory study.

Although substantial effort was made continually throughout the planning and testing of the intelligibility tests on the platforms to ensure technical rigour and robustness of results, not all aspects of the process could be controlled and concessions had to be made to allow other activities to occur at the station. Unexpected arrival of trains at the platforms, construction noise or other disturbances to the test subjects in the station could have skewed or reduced the repeatability and reliability of the results.

These factors would be expected to degrade the word score, thereby making a closer match between associated and measured STI values. However, as considerable care was taken to ensure robustness of the test data, and given that the word-score STIs are higher than measured, these factors do not seem to be the cause of the mismatch.

## 5 CONCLUSION

Measurements of the optimised public address (PA) system were undertaken during the commissioning phase of a project at an underground train station in Sydney, Australia to determine if the minimum STI requirement of 0.5 or CIS rating of 0.7 as per Australian Standard AS 1670-4:2018 was satisfied.

Initial measurements made on the platforms showed that the average STI was unlikely to meet the project's intelligibility requirements. As such, a word-score test using 1000-word phonetically-balanced words was proposed as an alternative method to demonstrate compliance through a conversion to a CIS or STI rating, using the conversion graph provided in AS1670.4:2018. The CIS and STI ratings from the word score test complied with the project's requirements, but this resulted in a mismatch between the direct STI measurements using STIPA signal and the STI results derived from word score testing.

An analysis of the individual word scores at test locations on the platform showed that architectural features and speaker settings affected the word score and STI ratings. A comparison of the measured STI and word score results with relationships in existing literature between these ratings was undertaken, and it was found that the measurements on the platform best matched the relationship from an Anderson and Kalb study. The reason for this is not understood, given that the study by Morales et al. was intended to address the weaknesses in the Anderson and Kalb study.

## 6 REFERENCES

1. 'Fire detection, warning, control and intercom systems - System design, installation and commissioning, Part 4: Emergency warning and intercom systems'. Standards Australia Limited, Dec. 21, 2018.
2. P. W. Barnett and R. D. Knight, 'The Common Intelligibility Scale', in *Proceedings of the Institute of Acoustics*, 1995, vol. 17 Part 7, pp. 199–204.
3. T. R. Letowski and A. A. Scharine, 'Correlational Analysis of Speech Intelligibility Tests and Metrics for Speech Transmission', ARL-TR-8227, Dec. 2017.
4. 'Method for Measuring the Intelligibility of Speech over Communication Systems'. Acoustical Society of America, 2020.

5. J. P. Egan, ‘Articulation testing methods’, *The Laryngoscope*, vol. 58, no. 9, pp. 955–991, Sep. 1948, doi: 10.1288/00005537-194809000-00002.

6. H. J. M. Steeneken and T. Houtgast, ‘A physical method for measuring speech-transmission quality’, *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, Jan. 1980, doi: 10.1121/1.384464.

7. *BSTJ 8: 4. October 1929: Articulation Testing Methods. (Fletcher, H.; Steinberg, J.C.)*. 1929. Accessed: Oct. 25, 2024. [Online]. Available: http://archive.org/details/bstj8-4-806

8. M. L. Hawley, R. Y. Litovsky, and J. F. Culling, ‘The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer’, *J. Acoust. Soc. Am.*, vol. 115, no. 2, pp. 833–843, Jan. 2004, doi: 10.1121/1.1639908.

9. A. K. Nábělek and P. K. Robinson, ‘Monaural and binaural speech perception in reverberation for listeners of various ages’, *J. Acoust. Soc. Am.*, vol. 71, no. 5, pp. 1242–1248, May 1982, doi: 10.1121/1.387773.

10. A. W. Bronkhorst and R. Plomp, ‘The effect of head-induced interaural time and level differences on speech intelligibility in noise’, *J. Acoust. Soc. Am.*, vol. 83, no. 4, pp. 1508–1516, Apr. 1988, doi: 10.1121/1.395906.

11. B. W. Anderson and J. T. Kalb, ‘English verification of the STI method for estimating speech intelligibility of a communications channel’, *J. Acoust. Soc. Am.*, vol. 81, no. 6, pp. 1982–1985, Jun. 1987, doi: 10.1121/1.394764.

12. L. Morales, S. Dance, B. Shield, and G. Leembruggen, ‘Speech Transmission Index for the English Language Verified Under Reverberant Conditions with Two Binaural Listening Methods: Real-Life and Headphones’, *J. Audio Eng. Soc.*, vol. 62, no. 7/8, pp. 493–504, Aug. 2014, doi: 10.17743/jaes.2014.0029.