INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

MOS vs. AB: Evaluating Text-to-Speech Systems Reliably Using Clustered Standard Errors

Joshua Camp, Tom Kenter, Lev Finkelstein, Rob Clark

Google
{joshcamp,tomkenter,finklev,rajclark}@google.com

Abstract

The quality of synthetic speech is typically evaluated using subjective listening tests. An underlying assumption is that these tests are reliable, i.e., running the test multiple times gives consistent results. A common approach to study reliability is a replication study. Existing studies focus primarily on Mean Opinion Score (MOS), and few consider the error bounds from the original test. In contrast, we present a replication study of both MOS and AB preference tests to answer two questions: (1) which of the two test types is more reliable for system comparison, and (2) for both test types, how reliable are the results with respect to their estimated standard error? We find that while AB tests are more reliable for system comparison, standard errors are underestimated for both test types. We show that these underestimates are partially due to broken independence assumptions, and suggest alternate methods of standard error estimation that account for dependencies among ratings.

Index Terms: speech synthesis evaluation, listening tests, reproducibility

1. Introduction

Despite renewed interest in objective evaluations [1], subjective listening tests continue to be the gold standard for text-to-speech (TTS) evaluation and are standard practice across the field [2]. Among the most widely used are Mean Opinion Score (MOS), preference/AB tests, and MUSHRA[3, 4]. While exact setups differ, these tests all try to measure the (possibly relative) intelligibility, naturalness, similarity, or expressiveness of synthetic speech, with naturalness being the most common. While naturalness is an under-specified term [5], it is generally understood to measure roughly how "human-like" the speech is.

A natural question that arises from MOS results is: if system A scores higher than system B on MOS, under what conditions can we conclude that system A is better than system B, and is this comparison as reliable as a direct AB test between these systems? We assume that the MOS test setups in question are generally comparable in terms of the platform and rating scale used for the evaluation, but not necessarily identical in terms of the individual listeners or when the test was run. Motivated by the question of system comparison and a suspicion that the confidence intervals we calculate for both MOS and AB are overly optimistic, we ask the following research questions:

- **RQ1**: Which of the two test types, MOS or AB, is more reliable for comparing TTS systems?
- **RQ2**: For both MOS and AB tests, how reliable are the results with respect to estimated standard errors?

To answer both of these questions, we present a replication study on MOS and AB listening tests. For RQ2 in particular, we use statistical methods that assume *clustering* is present in the data, i.e., that observations within clusters are not independent, but observations between clusters are. While these methods are not novel, to the best of our knowledge their application to TTS listening test data is. Lastly, the aim of the experiments presented is not to find the single best analysis technique – this will vary on a test-by-test basis. Rather, the aim is to demonstrate that cluster-based methods are both theoretically and empirically justified in the analysis of listening test data.

2. Related Work

A few studies investigate the reliability of MOS by repeating listening tests from past Blizzard Challenges[1]. [6] showed that while the correlation between historical and repeated MOS tests is high, old systems tend to score lower today compared to several years ago. In a similar study, [7] found the score drop between repeated historical evaluations to be particularly stark when evaluated alongside newer, higher-quality synthesis technologies, suggesting that absolute scores are not meaningful due to anchoring effects from the other systems in the test. However, they still found the rank order of the systems to be preserved, indicating that as a comparison tool MOS is reliable.

Less work has been done to investigate the reliability of MOS on shorter time scales. An exception is [8], who repeated MOS tests in three different locations around the same time and found high reliability, especially when considering the confidence intervals of the scores. However, their tests bear little resemblance to most modern TTS evaluations, as they tested audio codecs in a tightly controlled laboratory environment, as opposed to synthetic speech using online crowdsourcing tools.

The impact of the number of listeners and test utterances on listening test reliability has also been studied. In a re-analysis of the 2013 Blizzard Challenge evaluations, [9] concluded that listening tests should consist of a minimum of 150 total judgements from at least 30 listeners; any less resulted in poor sensitivity and reliability. In [10], a technique to correct for score bias arising from listener and utterance variation is proposed. While both studies confirm the influence of listener and utterance variation on MOS, only existing listening test data was analyzed, meaning results could not be confirmed across multiple runs of the same test.

2.1. Parametric vs. non-parametric statistical analysis

In this work we analyze listening tests using parametric statistical methods such as the *t*-test. We acknowledge that non-parametric methods are commonly used for listening test analyses, with [10] and [11] advocating for their use because Likert-

[1] https://www.synsig.org/index.php/Blizzard_Challenge

10.21437/Interspeech.2023-2014

style response formats produce ordinal, rather than interval data. Thus, the argument goes, it is not valid to compare means, compute standard deviations, or indeed run any kind of parametric test. However, this point is disputed by many in the statistical and measurement theory literature [12, 13, 14], and even if we were to accept this notion, other studies illustrate that many parametric methods in fact show considerable robustness in the face of violated assumptions [15, 16]. This is important because the ordinalist interpretation is limiting in terms of the available analysis techniques, the most important of which in the context of this work is the confidence interval, which is widely-reported on listening tests [17, 18, 19, 20]. Indeed, the ITU recommendation upon which MOS is based advocates for the use of confidence intervals (§A.4.5, [3]). Hence, we feel justified in our interpretation of the MOS and AB response formats as interval scales.

3. Comparing TTS Systems Reliably

In this section we describe our replication study, which is designed to compare the reliability of MOS and AB (RQ1). Starting with a proprietary dataset consisting of the previous year of listening tests, we consider cases where, for each side compared in an AB test, two MOS test runs exist, testing the exact same test set for both the A and B side of the test. We call these tests MOS_A and MOS_B respectively, and the population we sample from is the set of (AB, MOS_A, MOS_B) triplets. For all of these triplets we run two hypothesis tests. For AB, we run a one-sample *t*-test with the null hypothesis that the mean preference score is 0.0. For MOS_A and MOS_B we run a two-sample paired *t*-test with the null hypothesis that the means of MOS_A and MOS_B are equal. Both tests use an alpha level of 0.01.

In order to make a fair comparison between MOS and AB, our sampling scheme is designed to select a diverse set of triplets that is not biased in favor of one test type. Using the results of the hypothesis tests, we break the population into 4 strata and sample equally from each stratum: (1) both *t*-tests were statistically significant, (2-3) one of the two *t*-tests was statistically significant, and (4) neither *t*-test was significant. Then, within each stratum, we perform two rounds of systematic sampling [21], one using the value of the AB *t*-statistic to order tests and the other using the value of the MOS_A vs. MOS_B *t*-statistic.

To compare the reliability of AB vs MOS_A - MOS_B we consider the Pearson correlation coefficient (PCC), and the Spearman rank correlation coefficient (SRCC) between original and repeated runs of each test.

3.1. Estimating uncertainty

To address RQ2, we investigate methods for estimating the uncertainty of test scores, and introduce a method to evaluate the accuracy of these estimates.

Because the ratings from a listening test only constitute a sample of all possible ratings, the mean rating (the "score" of the test) is only a point estimate of "true" score, and will therefore contain some error. The standard error of the mean (SE) measures the uncertainty about how much sample means deviate from the true mean. In particular, it is the standard deviation of the mean's sampling distribution. The SE is used to compute two important quantities: confidence intervals and *t*-statistics.

For an independent sample of size n from a population with a standard deviation of σ, the standard error of the mean $\sigma_{\bar{x}}$ is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{1}$$

However, the population standard deviation is typically not known, so the SE cannot be calculated exactly and estimation techniques must be employed. Accurate estimations of the SE are critical to the reliability of listening tests – overestimation leads to Type II errors and underestimation leads to (arguably more serious) Type I errors. In this paper, we will compare four methods of standard error estimation to see which is most predictive of the replication study results.

3.1.1. Analytical method (AM)

The most common method, which we refer to as the *analytical method (AM)*, is to use the sample standard deviation as a point estimate for the population standard deviation σ in Eq. 1.

3.1.2. Standard bootstrapping (SB)

An alternative to the analytical method is *bootstrapping*; rather than estimating the population standard deviation, we estimate the standard error directly using the empirical distribution of our sample as an approximation of the population distribution [22]. The procedure is as follows: let $\mathbf{X} = X_1, ..., X_n$ be an independent random sample of size n. For $i = 1, ...k$, sample n items with replacement from $\mathbf{X}$, and compute the sample mean $\hat{\mu}_i$. For a sufficiently large k (we use 10,000), the set of means $\hat{\boldsymbol{\mu}} = \{\hat{\mu}_1, ...\hat{\mu}_k\}$ approximates the sampling distribution of the mean. The sample standard deviation of $\hat{\boldsymbol{\mu}}$ is the bootstrap estimate of the SE.

3.1.3. Cluster bootstrapping (CB)

The analytical method and standard bootstrapping assume observations are independent, which usually does not hold for listening tests. In our replication study, listeners rate multiple stimuli, which results in correlations between data points. We suppose that the set of ratings given by a single listener forms a *cluster* of ratings, such that items within a cluster are dependent, but items between clusters are independent.

A modification can be made to the bootstrapping procedure to account for clustered data. Suppose our sample is organized into m non-empty clusters $C_1, ..., C_m$, with each observation being assigned to exactly one cluster. To compute a bootstrap sample mean, $\hat{\mu}_i$, instead of sampling individual *observations*, we sample entire *clusters*. When each cluster contains an equal number of observations, we simply need to resample m clusters with replacement to arrive at a sample of size n. When the number of observations in each cluster are different, we resample clusters until we reach a cluster C_l that will increase the sample size to some $n' \geq n$. From C_l, we then randomly select $|C_l| - (n' - n)$ observations. This gives us a sample of size n and we can proceed with the procedure outlined in §3.1.2.

3.1.4. Effective sample size correction (ESS)

The final SE estimation method we investigate uses Eq. 1, but adjusts the sample size n to account for clustering in the sample. We use the *design effect* (D_eff), introduced by [23], which expresses the factor by which the sampling variance of a complex sampling design is inflated relative to the variance that would be observed using simple random sampling. Using this quantity, we can estimate the effective sample size, n_eff, as $\frac{n}{D_\text{eff}}$ (see [23] for full details regarding this method). The concept of the design effect is broadly applicable, but in our case, it allows us to take into account any clustering present in our sample.

There are multiple ways of calculating the design effect. We use the deff$_c$ formula from [24]. In this formula, we assume all sampling weights equal 1, and we estimate the intraclass correlation coefficient ρ using the ICC(1) method from [25], with the classes being the sets of ratings given by a single listener.

3.2. Evaluating standard error estimation

Because computing the true standard error requires the unknown population parameter σ, we evaluate the accuracy of our SE estimates indirectly using the following method.

For each listening test $i = 1, ..., n$, let m_i^1 and m_i^2 be the original and repeated test scores, respectively. We define the mean absolute difference (MAD) as:

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n} |m_i^1 - m_i^2| \tag{2}$$

Now let M_i^1 and M_i^2 be random variables representing the original and repeated test scores. Because all original and repeated pairs are run within 1 year of each other, we assume they are identically distributed, and because M_i^1 and M_i^2 represent sample means, we assume $M_i^1, M_i^2 \sim \mathcal{N}(\mu_i, \sigma_{i,\bar{x}}^2)$ by the Central Limit Theorem, where μ_i is the true mean score for test i and $\sigma_{i,\bar{x}}$ is the standard error. Assuming independence, $M_i^1 - M_i^2 \sim \mathcal{N}(0, 2\sigma_{i,\bar{x}}^2)$, and $E[|M_i^1 - M_i^2|] = \frac{2\sigma_{i,\bar{x}}}{\sqrt{\pi}}$ (half-normal distribution). We can then compute the mean *expected* absolute difference (MEAD) of original and repeated scores:

$$\text{MEAD} = \frac{1}{n}\sum_{i=1}^{n} E[|M_i^1 - M_i^2|] = \frac{1}{n}\sum_{i=1}^{n} \frac{2\sigma_{i,\bar{x}}}{\sqrt{\pi}} \tag{3}$$

We will compare the measured MEAD for each SE estimation technique to the observed MAD, with MEAD values closer to the MAD indicating the SE is more accurate *on average*.

4. Experimental Setup

The set of listening tests considered were crowd-sourced over a 1 year period from September 2021 to September 2022. We chose 1 year to give us a large enough pool to sample from, while being short enough that we do not expect scores to drop due to advancements in synthesis technology. For MOS, listeners are presented with a speech sample and are asked to rate how 'natural' it sounds on a 9-point scale from 1 to 5 with 0.5 increments, with whole number points being labelled *poor*, *bad*, *fair*, *good*, and *excellent*. For AB, listeners are presented with two speech samples and are asked which side sounds 'better' on a 7-point scale from -3 to +3 with 0 labelled *about the same* and the three points on both sides labelled *slightly better*, *better*, and *much better*.

In total, we sample 10 triplets per stratum (40 overall). This amounts to 113 listening tests in total, due to overlap between MOS$_\text{A}$ and MOS$_\text{B}$ tests between triplets. English varieties (US, UK, AU, IN) make up 60% of tests, with the remaining 40% coming from French, Portuguese, Chinese, Spanish, German, and Hindi.

The test parameters are the same across original and repeated tests: tests are run using the exact same questions and response formats on the same sets of approximately 1000 stimuli (μ: 966; σ: 101), with 1 rating per stimulus. The maximum number of stimuli allocated to one listener is held constant between the original and repeated runs, and it was possible for the same listener to participate in both tests. The repeated tests were run in October and November 2022.

Table 1: *Pearson correlation coefficient (PCC) and Spearman rank correlation coefficient (SRCC) per test type with 95% confidence intervals.*

Test Type	**PCC**	**SRCC**
MOS	0.956 ± 0.02	0.896 ± 0.05
AB	0.975 ± 0.02	0.809 ± 0.11
MOS$_\text{A}$- MOS$_\text{B}$	0.835 ± 0.10	0.596 ± 0.21
MOS$_A$- MOS$_{RandomB}$	*0.945 ± 0.04*	*0.893 ± 0.07*

Table 2: *Mean expected absolute difference (MEAD) for each standard error estimation technique. Values closer to MAD (top line) are better.*

		MOS	**AB**
MAD		*0.068*	*0.050*
MEAD – iid-based	AM	0.024	0.029
	SB	0.024	0.029
MEAD – cluster-based	CB	0.069	0.045
	ESS	**0.068**	**0.046**

5. Results and Analysis

5.1. MOS vs. AB

Table 1 shows the correlations between original and repeated tests. MOS shows high correlations, in line with the results reported by [6]. Furthermore, we found no evidence that scores decreased over the 1 year period, with the overall mean difference between original and repeated tests being 0.015 ± 0.021 (95% CI). This indicates that absolute scores are reasonably stable, providing the tests use the same setup and are run within a short enough time frame.

Which of the two test types, MOS or AB, is more reliable for comparing TTS systems (RQ1)? Table 1 shows that for the system comparisons in our experiments, a direct AB is more reliable than comparing separate MOS tests, though confidence intervals overlap for SRCC. The drop in correlation from MOS to MOS$_\text{A}$ - MOS$_\text{B}$ can be partially attributed to the choice of A and B – in our dataset the two systems being compared tend to be similar in quality, meaning scores are concentrated around 0. To illustrate this, we run a version of MOS$_\text{A}$ - MOS$_\text{B}$, MOS$_\text{A}$ - MOS$_\text{RandomB}$, where each side A is paired with a random side B, and find the correlations to be much higher than MOS$_\text{A}$ - MOS$_\text{B}$. Therefore, we conclude that while MOS exhibits acceptable reliability, AB tests should be preferred over separate MOS tests for system comparison, especially when the systems under test are of a similar quality.

5.2. Accuracy of standard error estimates

While Table 1 shows reliability in terms of raw scores, we are also interested in how repeated tests compare in terms of their estimated standard errors (RQ2). Table 2 shows the MAD for MOS and AB, along with the MEAD scores for each standard error estimation method. We see that iid-based methods systematically underestimate the MAD by a factor of 1.67 for AB and 2.83 for MOS. Both cluster-based methods show large improvements, nearly matching the observed MAD.

Table 2 shows that cluster-based SE estimation methods produce more accurate SEs *on average*, but do they capture anything meaningful at the individual test level? We test this by checking if tests with higher SE estimates tend to have larger differences between the original and repeated tests. Table 3 shows the PCC and SRCC between the estimated SE and the

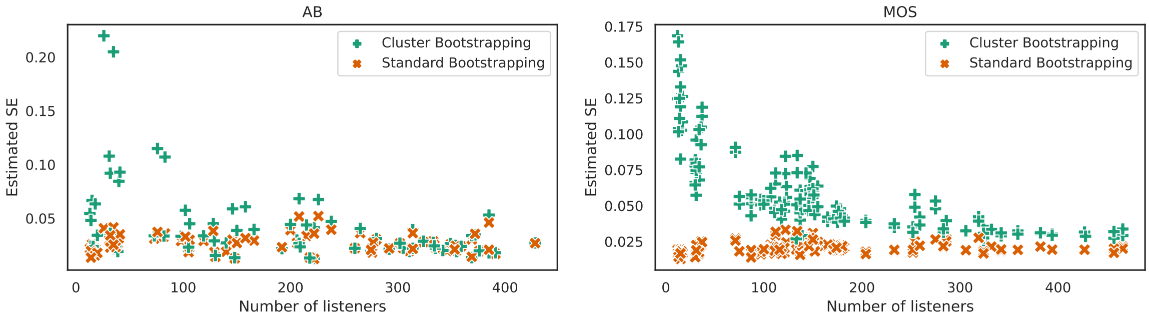

Figure 1: *SE estimates for both bootstrapping methods vs number of listeners for AB (left) and MOS (right).*

Table 3: *Correlations (PCC and SRCC) between absolute differences between original and repeated scores and SE estimates with 95% confidence intervals.*

Test	SE estimator	PCC	SRCC
MOS	AM	0.17 ± 0.22	0.07 ± 0.23
	SB	0.16 ± 0.22	0.06 ± 0.23
	CB	0.28 ± 0.21	0.38 ± 0.20
	ESS	0.26 ± 0.21	0.38 ± 0.20
AB	AM	0.56 ± 0.22	0.59 ± 0.21
	SB	0.56 ± 0.22	0.59 ± 0.21
	CB	0.64 ± 0.19	0.45 ± 0.25
	ESS	0.62 ± 0.20	0.48 ± 0.25

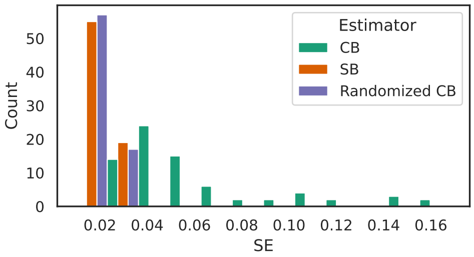

Figure 2: *SE estimates for MOS tests for SB, CB, and Randomized CB.*

absolute difference between original and repeated tests for the four SE estimation strategies. The results for MOS are clear: cluster-based methods correlate with test differences and iid-based methods do not, indicating that the methods are doing more than simply scaling up iid-based SEs. However, there seems to be little to no effect on AB tests. This discrepancy may be explained by the effect of the number of listeners on both SE and score differences, which we discuss in the next section.

5.3. Effect of number of listeners

As the CB method clusters at listener level (§3.1.3), we are interested in the relationship between the number of listeners and the estimated SE for iid-based and cluster-based methods. Figure 1 shows the results for SB and CB (results for AM and ESS were similar, but were omitted for clarity). Cluster-based SEs correlate strongly with the number of listeners, roughly exhibiting a power law relationship. The effect is less pronounced for AB tests, indicating they are less affected by the number of listeners than MOS. We hypothesize this is due to the comparative nature of the task, which makes the rating scale more interpretable than MOS, thus leading to higher rater agreement.

SE estimates from CB correlate with the number of listeners, but could this simply be due to the fact that the resampling units (clusters) are larger the fewer listeners there are? To analyze this, we run a version of CB where we keep the number and size of clusters from normal CB, but rather than each cluster corresponding to one listener's ratings, we randomly allocate ratings to clusters. Figure 2 shows the histogram of Randomized CB SEs alongside SB and CB for all MOS tests. Randomized CB's estimates are very close to SB, indicating that the inflated SEs from CB are due to genuine dependencies within clusters and not larger sampling units.

Finally, we investigate whether the number of listeners who took part in a test correlates with the score differences between original and repeated tests. We found the correlation to be statistically significant ($p < 0.01$) for MOS, and not for AB. This points to why cluster-based SEs were only predictive of test differences for MOS; SE and MOS differences are *both* correlated with the number of listeners, while for AB only SE was correlated with the number of listeners.

6. Conclusions and Future Work

We presented a replication study on MOS and AB listening tests run within a 1 year period. We found AB to be more reliable than MOS for system comparison. However, for both MOS and AB, we found that out-of-the-box approaches to standard error estimation resulted in underestimates up to a factor of 2.8. We showed that this underestimation can partially be attributed to broken independence assumptions in the listening test data that arise from the same listener rating multiple stimuli. We showed that *cluster-based* methods produced better standard error estimates, and recommend their adoption for listening test analyses. We also found that AB reliability was less influenced by the number of listeners, suggesting that comparative tests may be more appropriate for situations in which a large number of listeners cannot be recruited.

In this study we investigated MOS tests where only stimuli from one system were rated, and each stimulus was rated only once. Future work should investigate more complex setups involving multiple systems such as MOS using Latin square designs or MUSHRA, as these may prove to be more reliable. When each stimulus is rated multiple times, additional dependencies are introduced because scores within a single item will be correlated. For this case, future work could investigate mixed-effects models, which seem promising for their ability to model *crossed* random effects [26], the random effects being listeners and stimuli.

7. References

[1] W. C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," in *Proc. Interspeech 2022*, 2022, pp. 4536–4540.

[2] Z. Ling, X. Zhou, and S. King, "The Blizzard Challenge 2021," in *Proc. Blizzard Challenge Workshop*, 2021.

[3] I. Rec, "P. 800: Methods for subjective determination of transmission quality," *International Telecommunication Union, Geneva*, vol. 22, 1996.

[4] ITU-R BS. 1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," *International Telecommunication Union*, 2003.

[5] R. Dall, J. Yamagishi, and S. King, "Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation," in *Proc. Speech Prosody 2014*, 2014, pp. 1012–1016.

[6] E. Cooper and J. Yamagishi, "How do Voices from Past Speech Synthesis Challenges Compare Today?" in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 183–188.

[7] S. Le Maguer, S. King, and N. Harte, "Back to the Future: Extending the Blizzard Challenge 2013," in *Proc. Interspeech 2022*, 2022, pp. 2378–2382.

[8] J. Holub, H. Avetisyan, and S. Isabelle, "Subjective speech quality measurement repeatability: comparison of laboratory test results," *International Journal of Speech Technology*, vol. 20, pp. 69–74, 2017.

[9] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? No! — An empirically-supported critique of Interspeech 2014 TTS evaluations," in *Proc. Interspeech 2015*, 2015, pp. 3476–3480.

[10] A. Rosenberg and B. Ramabhadran, "Bias and Statistical Significance in Evaluating Speech Synthesis with Mean Opinion Scores," in *Proc. Interspeech 2017*, 2017, pp. 3976–3980.

[11] R. A. J. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," *Proc. BLZ3-2007 (in Proc. SSW6)*, 2007.

[12] J. Gaito, "Measurement scales and statistics: Resurgence of an old misconception," *Psychological Bulletin*, vol. 87, no. 3, pp. 564–567, 1980.

[13] S. E. Harpe, "How to analyze Likert and other rating scale data," *Currents in Pharmacy Teaching and Learning*, vol. 7, no. 6, pp. 836–850, 2015.

[14] G. Norman, "Likert scales, levels of measurement and the "laws" of statistics," *Advances in Health Sciences Education*, vol. 15, pp. 625–632, 2010.

[15] G. V. Glass, P. D. Peckham, and J. R. Sanders, "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance," *Review of Educational Research*, vol. 42, no. 3, pp. 237–288, 1972.

[16] A. J. Vickers, "Comparison of an ordinal and a continuous outcome measure of muscle soreness," *International Journal of Technology Assessment in Health Care*, vol. 15, no. 4, pp. 709–716, 1999.

[17] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, Robust and Controllable Text to Speech," in *Proc. NeurIPS*, 2019, pp. 3171–3180.

[18] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," in *Proc. Interspeech 2020*, 2020, pp. 4432–4436.

[19] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. J. Weiss, and Y. Wu, "Parallel Tacotron: Non-Autoregressive and Controllable TTS," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5709–5713.

[20] H. Cho, W. Jung, J. Lee, and S. H. Woo, "SANE-TTS: Stable And Natural End-to-End Multilingual Text-to-Speech," in *Proc. Interspeech 2022*, 2022, pp. 1–5.

[21] W. G. Madow and L. H. Madow, "On the theory of systematic sampling, I," *The Annals of Mathematical Statistics*, vol. 15, no. 1, pp. 1–24, 1944.

[22] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. CRC press, 1994.

[23] L. Kish, *Survey Sampling*. Wiley, 1965.

[24] P. Lynn and S. Gabler, "Approximations to b* in the Prediction of Design Effects Due to Clustering," *Survey Methodology*, vol. 31, pp. 101–104, 2005.

[25] D. Liljequist, B. Elfving, and K. Skavberg Roaldsen, "Intraclass correlation – A discussion and demonstration of basic features," *PloS ONE*, vol. 14, no. 7, p. e0219854, 2019.

[26] R. H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *Journal of Memory and Language*, vol. 59, no. 4, pp. 390–412, 2008.