**Comparing the Accuracy of ACE and WER Caption Metrics When Applied to Live Television Captioning**

Tian Wells, Dylan Christoffels, Christian Vogler, and Raja Kushalnagar

[t.wells@columbia.edu](mailto:t.wells@columbia.edu), [dylan.christoffels@gallaudet.edu](mailto:dylan.christoffels@gallaudet.edu), [christian.vogler@gallaudet.edu](mailto:christian.vogler@gallaudet.edu), [raja.kushalnagar@gallaudet.edu](mailto:raja.kushalnagar@gallaudet.edu)

Gallaudet University, 800 Florida Ave NE, Washington, DC 20002

**ABSTRACT**

The development of caption metrics is a new and ongoing growth in the accessibility research community. However, little work has been done comparing the effectiveness of newly developed caption metrics. More specifically, in low accuracy settings such as live television, where users report the most difficulty using captions. Through a user study with fifteen participants, we compared two caption metrics systems, Word Error Rate (WER) and Automated-Caption Evaluation (ACE), for their accuracy in evaluating caption quality in live television. We compared human-perceived quality statistics with each caption metric’s data. Analysis of the correlation between human statistics and each caption metric found that WER had a slightly higher correlation with participants. We found that ACE was more sensitive to errors that WER, and penalized captions more than participants. However, the difference in performance between WER and ACE was not statistically significant, and neither WER nor ACE are optimized for use with live television captioning. Future work should explore how caption metrics could be better optimized for use with live television.

**Keywords**

Word Error Rate (WER), Automated-Caption Evaluation (ACE), Deaf or Hard of Hearing (DHH), Caption Metrics, Accessibility, Live Television

1. **INTRODUCTION**

The task of evaluating the quality of captions is not an easy one. Quality in captioning has no clear definition, but for deaf and hard of hearing (DHH) individuals that use captions, quality is often synonymous with comprehension. Real-time captioning used for news, for example, are a lifeline for DHH individuals that wish to stay up to date with current events. Additionally, live videos on social media platforms such as Facebook and Instagram are a widespread tool DHH individuals use to interact with friends. The quality of captions goes hand in hand with their ability to relay information accurately and effectively to DHH viewers.

However, a lack of resources and cost restrictions often result in mistakes during the captioning process, particularly in live captioning. A 1999 study found a 17% word error rate in real-time captions used across broadcast news [1]. Because computer systems used to evaluate captions employ a universal dictionary, live captions also often will not include proper nouns such as names, places, and people in the broadcast [2]. Generally, captions generated with automatic speech recognition (ASR) are limited by the capability of today’s technology and are prone to translation errors. Captions created with respeaking, when a trained individual repeats what is heard into a voice recognition software, are also prone to user-based errors, as well as software errors. Similarly, captions generated with steno captioning, a process that often requires a skilled Court Reporter to create real-time captions, leads to user-based errors [3].

Stenocaptioning is the most common captioning method in live television. The software used in this process is able to report errors automatically but only works by determining whether keystrokes match an English word [3]. The software does not account for mistakes, such as incorrect word substitution, and also falsely identifies adequate rephrasing as mistakes. Consequently, holistic assessment of the quality of stenography-generated captions will require other methods—caption metrics.

In order to assess the quality of captions, caption metrics numerically compare what is actually spoken with the captions. Common metrics include Word Error Rate (WER), Weighted Word Error Rate (WWER), Automated-Caption Evaluation (ACE), and NER, which are helpful for DHH individuals in rating the quality of a captioned video [4]. Although there are many different caption metrics, there is little to no research exploring how these metrics compare to each other. In a 2017 study, researchers compared the efficacy of WER and ACE metrics, but the study design was limited in that it was never tested using real-time captioning. Instead, it used staged scenes with ASR-generated captions. Regardless, ACE was found to have a better correlation with participants’ experiences [6]. There is little research into this topic beyond this study, and there are many disparities in the state of caption quality assessment today for DHH individuals. Our aim is to compare the WER, WWER, and ACE captioning metrics and assess how well they correlate with quality.

**1.1 WER**

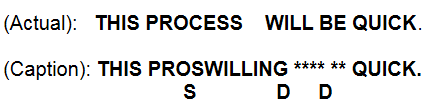
WER tests the performance of vocabulary continuous speech recognition in captions, comparing what is spoken on the TV with the captions displayed on screen [4]. WER is a metric that penalizes words that are incorrect, due to a mistake in one of the three categories: [3]

1. Substitution (S)  
   Substitution errors are where an erroneous word is substituted for the correct word in the reference transcript.
2. Deletion (D)  
   Deletion errors are when a word has been deleted or omitted from the transcript.
3. Insertion (I)   
   Insertion errors are when one word has been inserted into the transcript that was not spoken.

These three error categories do not have weights based on error severity. Errors add to a fully accurate score of 0%. The WER is calculated by the following formula.

**Figure 1.** The WER formula. Where S indicates substitution, D indicates deletion, I indicates insertion, and N indicates the total number of words in the original transcript.

In the equation above, the net amount of errors are divided by N - the total number of words in the reference text [3].



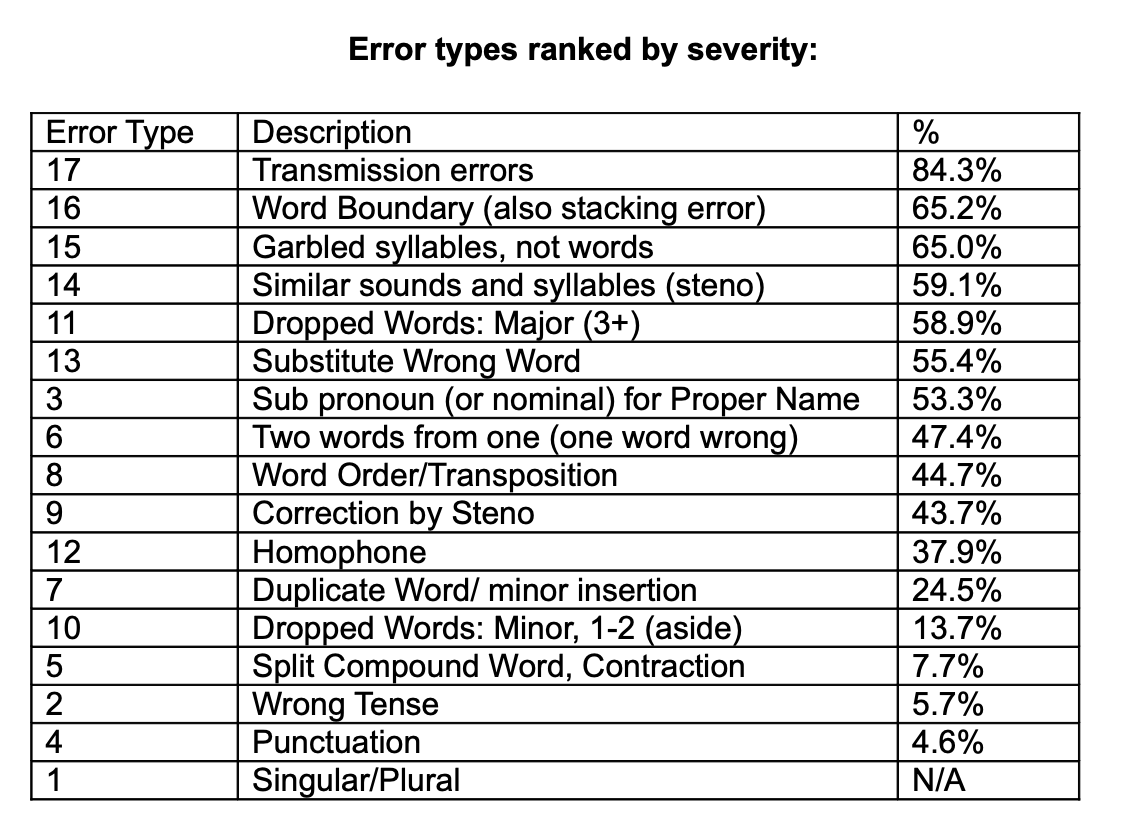
**Figure 2.** The image shows two different sentences. (Actual) is the spoken. (Caption) is the subtitling shown on the screen. The letters S and D stand for Substitution and Deletion. The stars show missing words, which are deletions. Using the formula above, there would be three errors out of 5 total words, giving a 60% Word Error Rate.

**1.2 WWER**

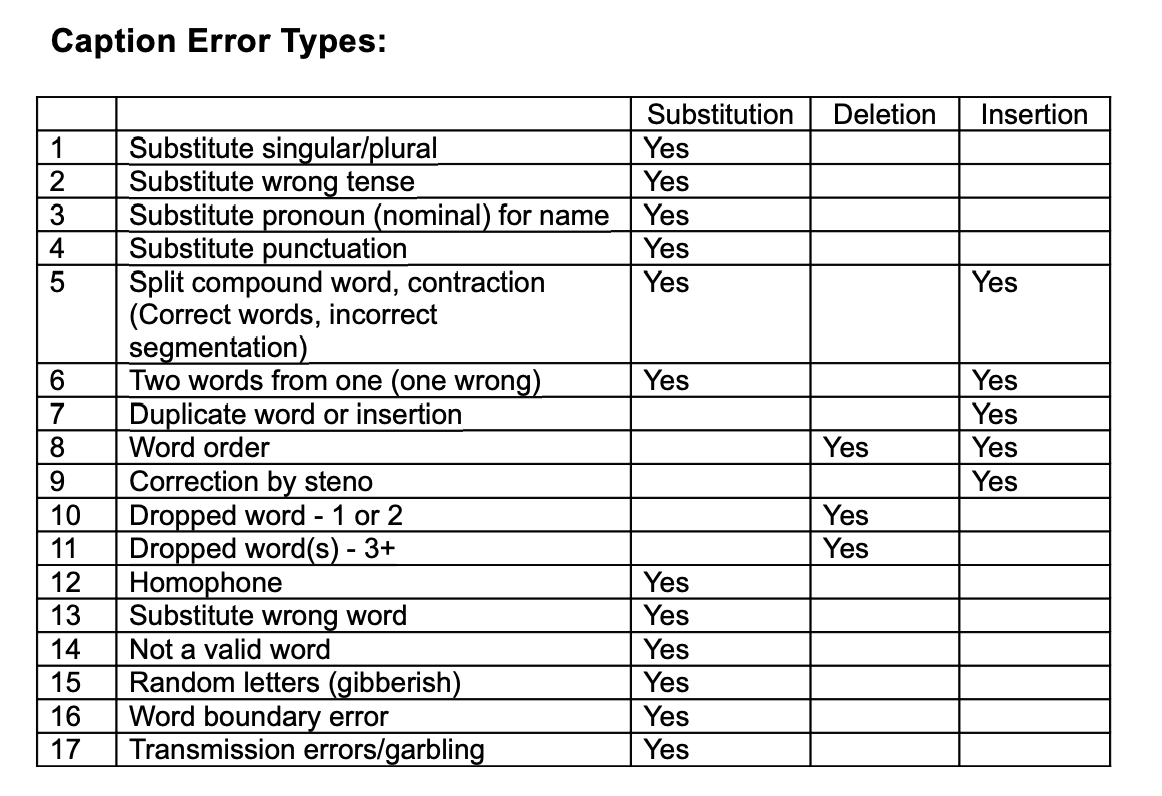
WWER is based on WER, except errors are instead weighted based on severity. There are 17 error types that each have a weight. Each of the 17 types belongs to one of the 3 WER error types. The calculation for WWER takes the sum of each error type multiplied by its weight and divides the sum by N, the number of words in the reference transcript [5].

**Figure 3.** The equation shows steps on how to calculate the WWER.

Below are the 17 different error types used for WWER. Generally, 1 is the least severe and 17 is the most.

****

**Figure 4.** The seventeen WWER Caption Error Types and which category each falls into: Substitution, Deletion, or Insertion. [5]

****

**Figure 5.** The results of the 2010 WGBH NCAM study. For each error type, the number expresses the percentage of respondents that determined the error type would greatly affect their understanding of the content. [5]

We will be adhering to the results of a 2010 WGBH National Center for Accessible Media (NCAM) study that surveyed DHH and hearing individuals to define the severity of each error type [3]. Shown above are the results. The table shows the percentage of respondents who thought that the caption error would “greatly affect” or “completely destroy my understanding” of the content [3]. The weight of each error is taken to be the decimal value of each result. For example, error type 17 will have a weight of 0.843.

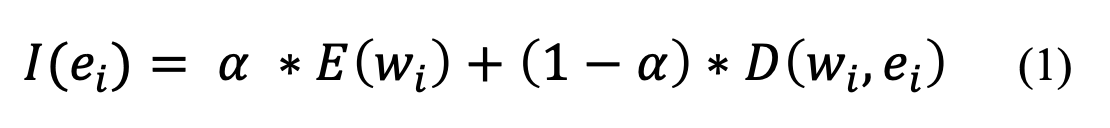
**1.3 ACE**

The Automated-Caption Evaluation (ACE) model is an automatic caption evaluation metric that was designed for use specifically with DHH individuals and ASR-generated captions. Unlike WER, it distinguishes between harmful and less harmful errors. ACE uses a word predictability score to measure keywords in a text and a semantic distance model (between the word actually spoken and the word displayed in a caption text) to approximate deviation [6].

Word predictability is calculated through assigning an entropy score to error words and their corresponding correct word. A higher entropy score corresponds to a more unpredictable error word. The entropy score of an insertion error word is calculated by comparing with the two adjacent words to the insertion error. Then, the word predictability score is calculated from the entropy score [6].

Semantic distance is calculated through a tool from Google called Word2vec [7]. Word2vec assigns a high score to two words that are largely different, such as “truthful” and “false”. It will assign a lower score to two words with more semantic similarity, such as “trust” and “trusted,” and in cases of insertion and deletion errors, the semantic deviation is scored based on the length of the inserted or deleted word [6].

Finally, the word predictability and semantic distance scores are combined using a weight of 0.65 (represented by 𝛼 in the metric formula). This indicates that the ACE score depends slightly more on the word predictability score than the semantic distance score. These scores are used to calculate the impact of an error with the following formula:



**Figure 6.** The equation shows steps on how to calculate the error by the ACE system. [6]

is the error word, and is the reference word in the caption text. 𝛼 is the experimentally determined weight, 0.65, that is applied to the scores. refers to the predictability score, and refers to the semantic distance score [6].

We will be using the ACE2 automatic evaluation tool for the purposes of this study.

**2. STUDY PREPARATION**

In order to assess how well the caption metrics assess quality in captioning, we compared each of the metrics’ scores alongside participants’ perceived quality scores of the captions. We used clips from live television with captions that have been generated through either ASR or steno captioning as stimuli for this study.

**2.1 Stimuli Selection and Preparation**

The stimuli used in this study were provided to us by the Gallaudet University REU AICT program. We were given access to over 200 recordings of segments of US and British live television, ranging from 20 minutes to over an hour in length. In order to curate a set of recordings that could be used as stimuli for our study, we selected 10 clips that would represent a variety of caption errors. We classified the caption errors in the TV recordings in correspondence with the 17 WWER error types, and selected 30 second- to 1 minute-long clips of the recordings that included the most diverse set of WWER errors. In doing so, we aimed to analyze the performances of each caption metric with almost all of the different caption error types. Recordings were also selected with their transmission error rate in mind. Only recordings with an SRT error rate of <1% were selected, in order to filter out errors that are caused by transmission errors rather than caption errors. Additionally, the content of the recordings were taken into account. News with themes relating to politics were generally avoided, as well as themes of violence and other potentially distressing content were filtered out. News segments relating to weather, sports, and natural catastrophes were primarily chosen to prioritize the comfortability of participants and rule out political or trauma-related biases. As a result, stimuli were clips taken from live segments of ABC 7 News, BBC World News, and CNN News.

After selecting the stimuli clips, accurate transcripts of the clips were obtained from the online Rev Captioning service. All descriptive and sensory details included in the accurate transcripts were removed. Then, we had to alter several aspects of the original error clips to be identical with the accurate captions from Rev. We made copies of the accurate transcripts, and edited the copies to match the errors of the original captions. These served as our error or “original” caption clip stimuli. This is because there were many disparities in captioning style, timing, and delay, between the error and accurate captions, so creating a copy with the correct caption style, timing, and delay ensured consistency with each playthrough.



**Figure 7.** A screenshot of Clip 2 (ABC News - Feb 5) with error captions.

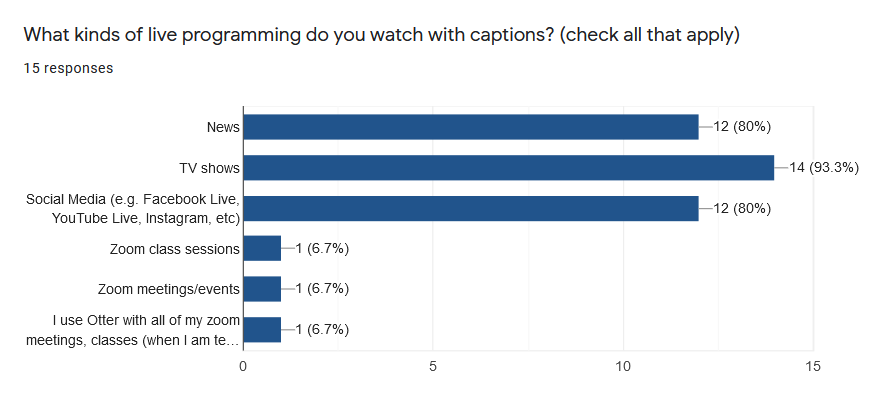
****

**Figure 8.** A screenshot of Clip 2 (ABC News - Feb 5) with accurate captions.

**2.2 Participant Recruitment**

We selected hearing, hard of hearing, and deaf participants that were 18 years of age or older, in order to include a variety of caption-users in the study. Our criteria was that participants had to have had experience with using captions for live television, ensuring they would be able to adequately read and notice errors in the captions from our stimuli. For the same reason, participants were required to have normal vision or corrected to normal vision with glasses or contacts. Participants were recruited from the Gallaudet University and Columbia University student and alumni communities through word of mouth and social media flyers.

Fifteen people were tested in total. Six deaf, six hearing, and three hard-of-hearing people participated in the study. The age of participants varied but the majority of participants were ages 18-24, making up 66.7% of the test population. The gender identities of participants were 66.7% female and 33.3% male. The common methods of communication amongst participants also varied. 66.7% of participants reported that they used spoken English most often and 33.3% reported that they primarily used American Sign Language. The instructions for the study were given in the participant’s preferred language.



**Figure 9.** The horizontal bar graph shows the percentage of respondents that would use captions with each type of programming.

**3. STUDY PROCEDURE**

Once the participants were gathered, we double-checked that they met the study’s criteria and then we conducted each individual study through Zoom. In the Zoom meeting, participants were asked to fill out two surveys. The first was for the main experiment, which included the 10 sets of stimuli clips, each set with error and accurate caption stimuli and an identical set of questions corresponding to each set. Participants were told to watch each clip one time through with captions and to respond to questions in the order they were shown. After being shown the first clip with caption errors, participants were asked if they noticed any caption errors and to rate the quality of the captions on a scale of 1 to 7 (poor to excellent). After being shown the second clip with accurate captions, participants were asked again to rate the quality of the original error captions on a scale of 1 to 7. For the final two questions, participants were asked if they felt they were able to fully understand the content of the first clip and were asked to compare the qualities of the error and accurate captions on a scale of 1 to 7 (little difference to big difference). Participants independently repeated this process 10 times for each set of stimuli and were given support and additional instructions over Zoom if needed.

**4. RESULTS AND ANALYSIS**

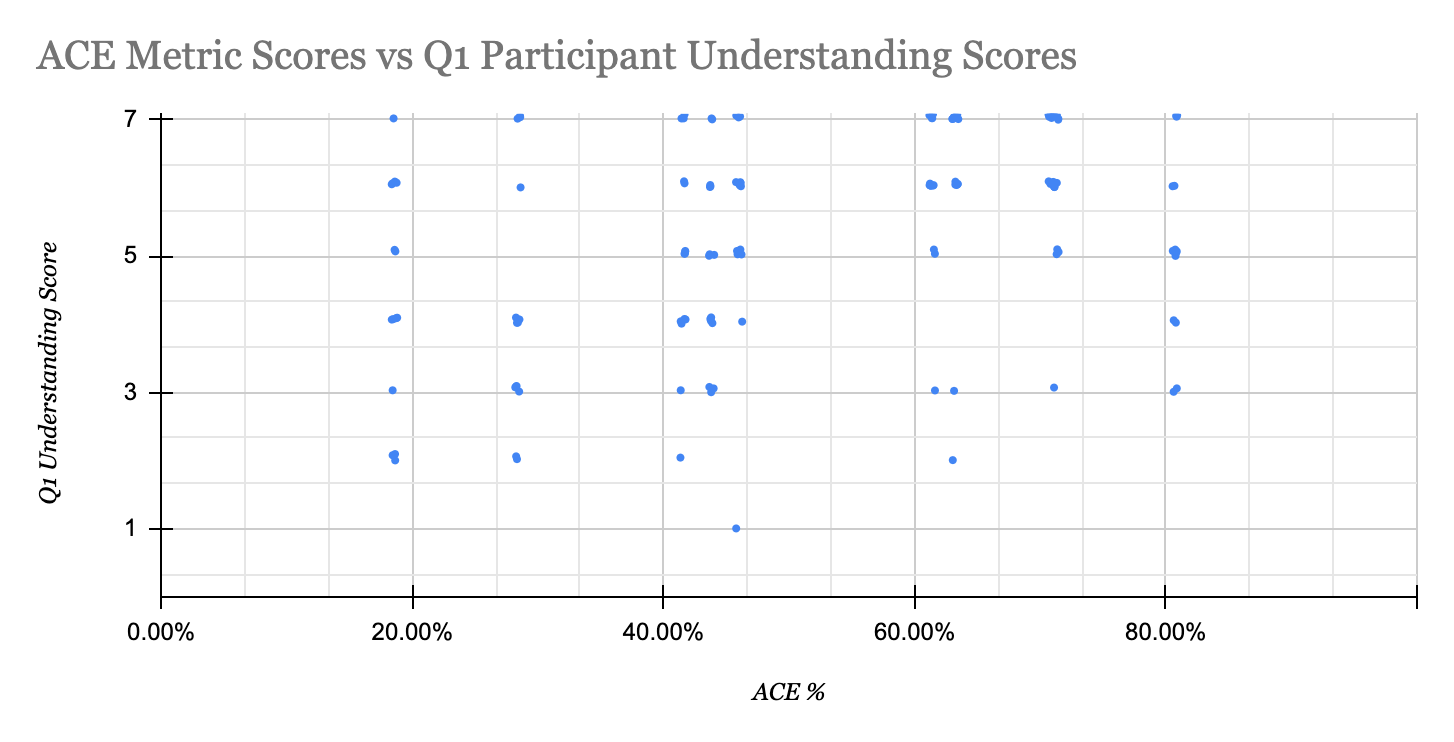
All data was obtained through Google forms. The data from the last two questions of each set of clips were chosen for analysis, as they best encapsulated caption quality and were asked after participants had watched both the error and accurate stimuli clips. We determined that participants would be able to best identify caption errors, and thus, caption quality, if they were able to compare the original caption errors with a correct transcript. The first question (**Q1**) chosen for analysis asks: “Do you feel you were able to fully understand the content of the error clip?”. The second question (**Q2**) chosen asks: “Comparing the original captions and the accurate captions, how would you rate the difference between the two?”

On average, participants reported that on a scale of 1 to 7, they were able to fully understand the content of the error clip at a 5.37 rating. Additionally, respondents had an average score of 3.91 when asked to compare the difference between the original and accurate captions on a scale of 1 to 7 (little difference to big difference).

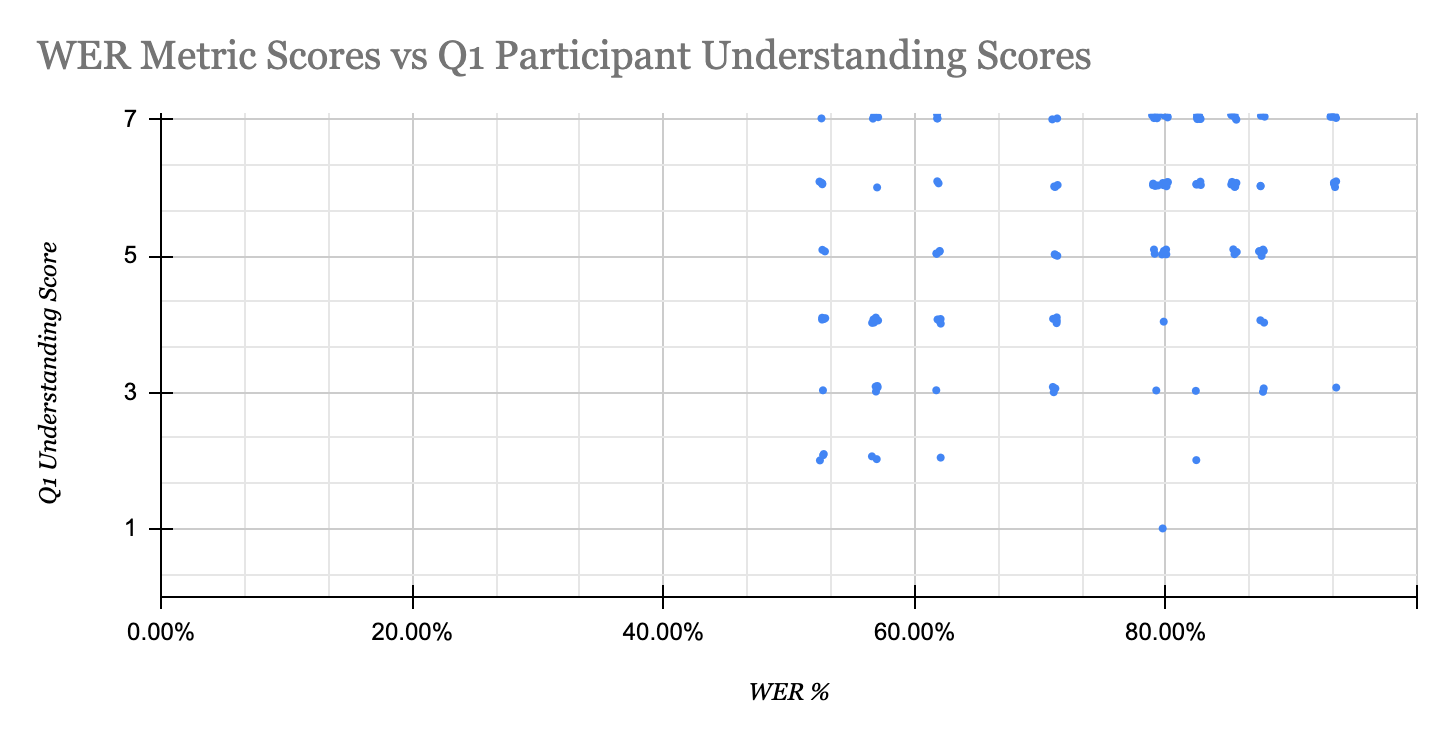
| Metric | Clip 1 | Clip 2 | Clip 3 | Clip 4 | Clip 5 | Clip 6 | Clip 7 | Clip 8 | Clip 9 | Clip 10 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ACE | 63.49% | 41.84% | 46.28% | 44.08% | 81.02% | 28.65% | 18.85% | 71.50% | 61.65% | 71.16% |
| WER | 82.80% | 62.10% | 80.20% | 71.40% | 87.90% | 57.10% | 52.90% | 85.70% | 79.40% | 93.60% |

**Figure 10.** The table shows the metrics data from ACE and WER for each stimuli clip, where higher percentages indicate higher accuracy.

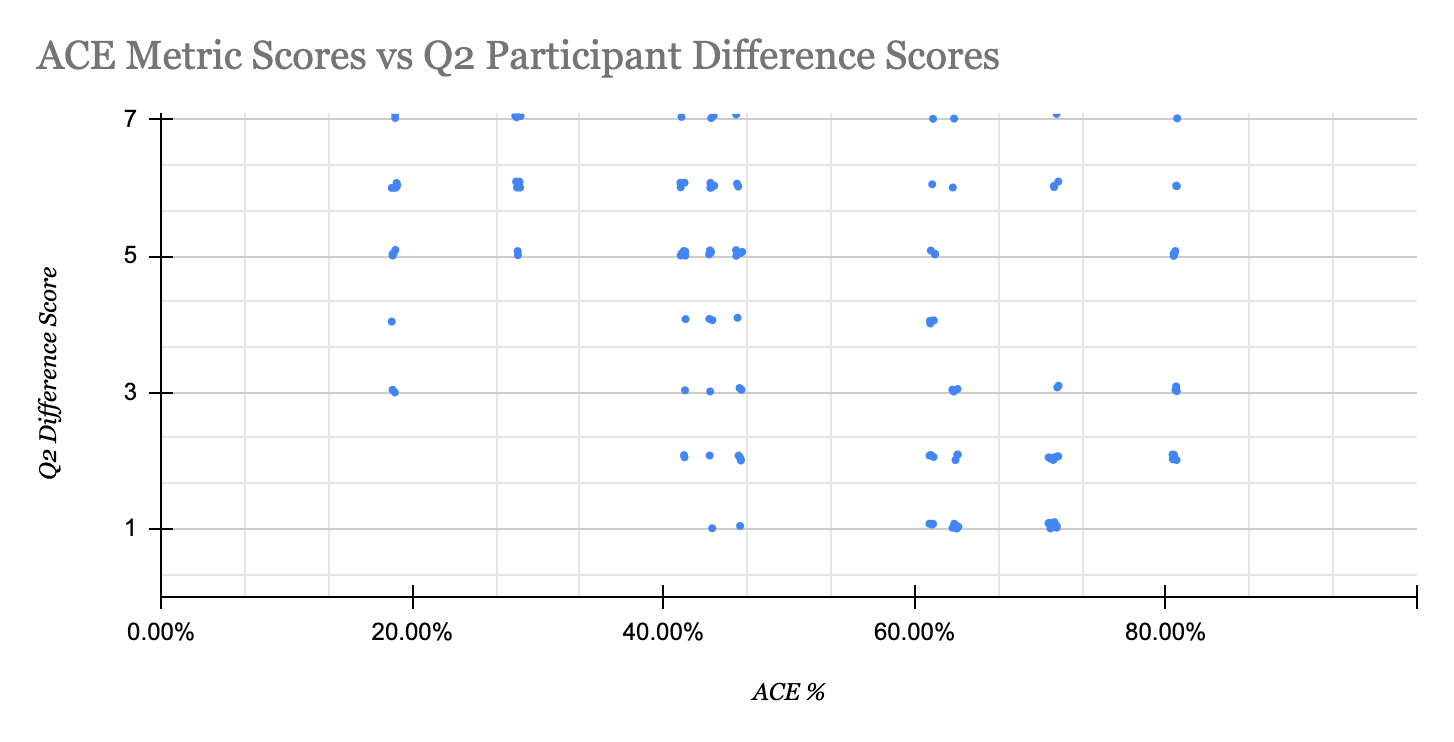
Figures 11-14 show the distribution of Q1 and Q2 scores in comparison to ACE and WER. Q1 is scored such that 1 indicates little to no understanding of the error captions, while 7 indicates complete understanding. As such, a positive relationship between responses and metric scores is expected. Q2 is scored such that 1 indicates little to no difference between the error and accurate captions, while 7 indicates a large difference. In this case a negative relationship between responses and metric scores is expected.



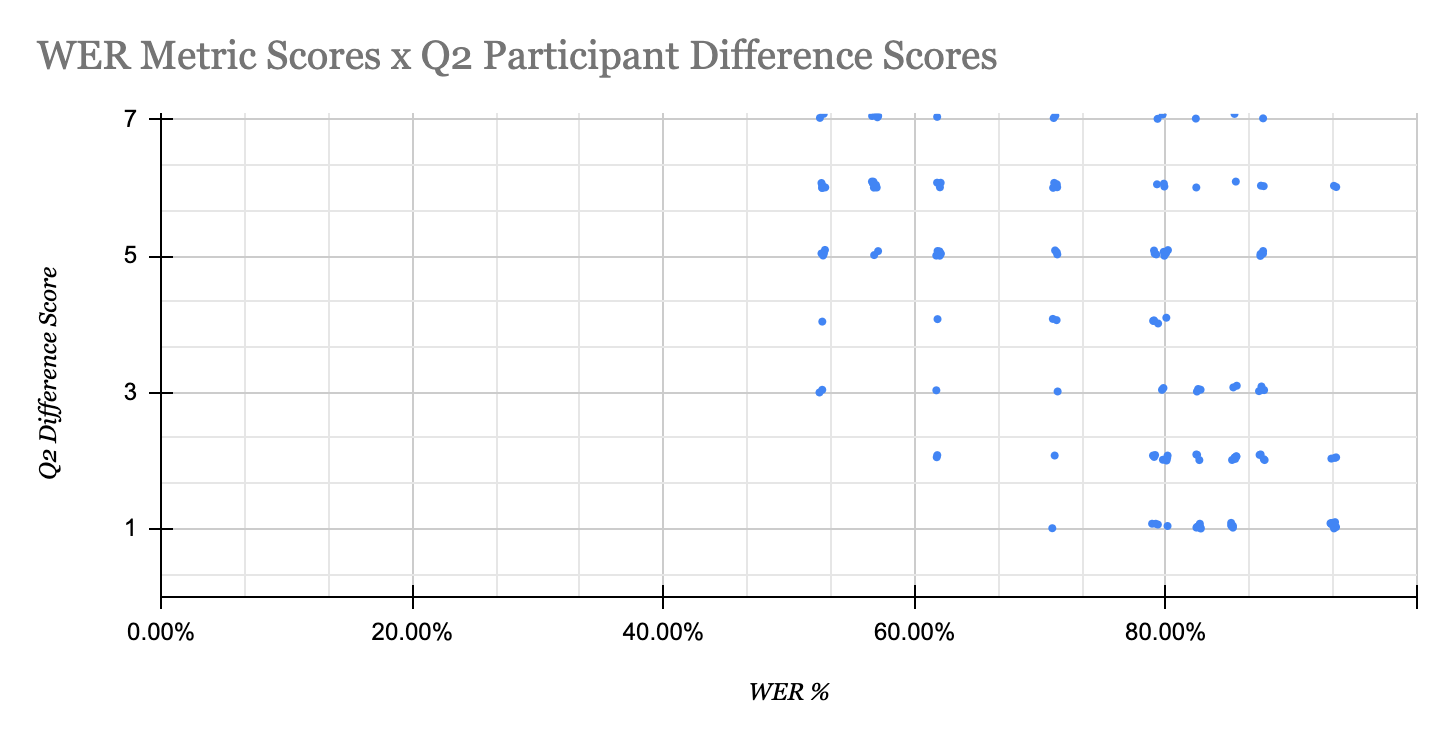
**Figure 11.** Scatter plot of all participant responses to Q1 for all stimuli with the corresponding ACE metric.



**Figure 12.** Scatter plot of all participant responses to Q1 for all stimuli with the corresponding WER metric.



**Figure 13.** Scatter plot of all participant responses to Q2 for all stimuli with the corresponding ACE metric.

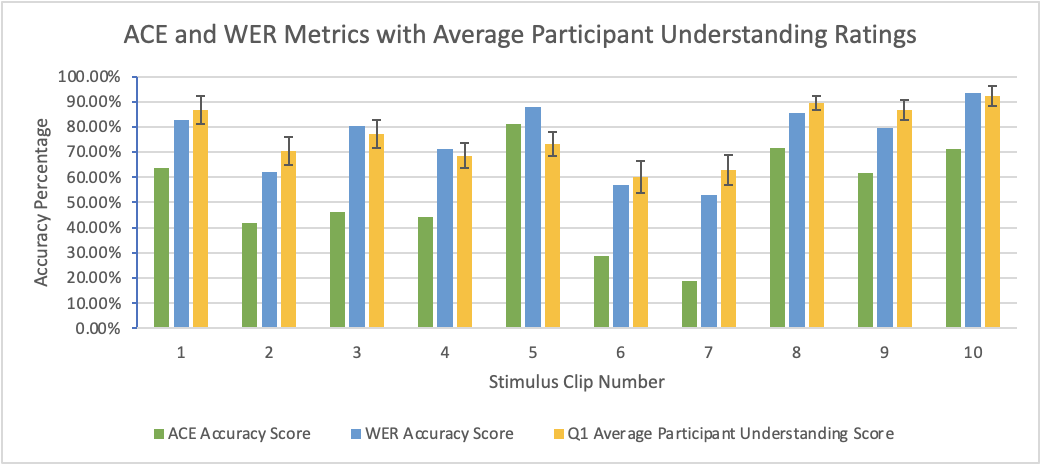


**Figure 14.** Scatter plot of all participant responses to Q2 for all stimuli with the corresponding WER metric.

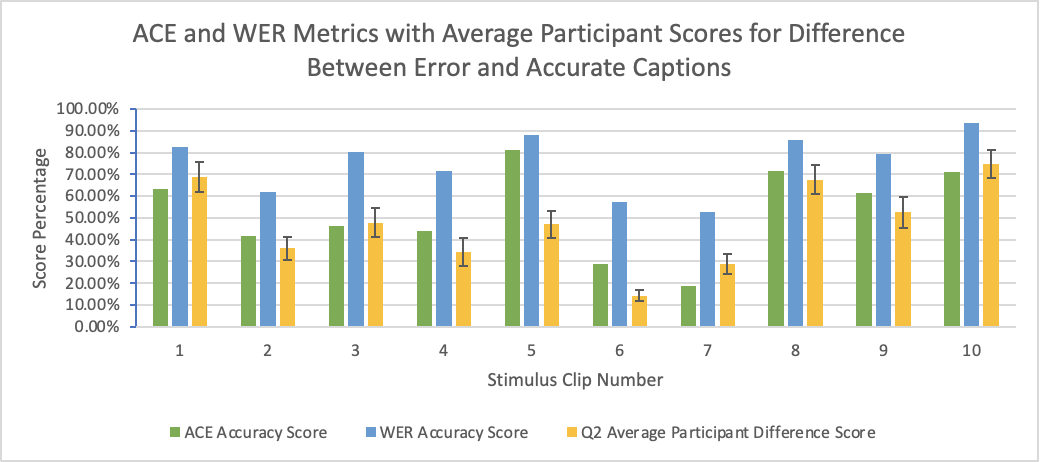
**4.1 Statistical Analysis**

As shown in Figure 10, the ACE and WER scores differ. For ACE, Clip #5 has the greatest accuracy and Clip #7 the lowest. For WER, Clip #10 has the greatest accuracy and Clip #7 the lowest. The WER scores are noticeably higher than ACE across all stimuli. Differences in ACE and WER can be attributed to their vastly different scoring methodologies, as mentioned in sections 1.1 and 1.3. Primarily we suspect ACE’s inclusion of word weights, and WER’s lack thereof. ACE more precisely scores caption errors, which also contributes to its lower accuracy scores overall, since errors are more penalized as a result. It is also important to note that ACE was designed specifically with deaf and hard of hearing people in mind, while our study includes deaf, hard of hearing, and hearing people.

Figure 15 displays the ACE and WER metric scores and the average participant scores for Q1 for each of the 10 sets of clips. Responses to Q1 were averaged and converted to a percentage. Similarly, in Figure 16, responses to Q2 were averaged, converted to a percentage, and inverted, such that a score of 100% indicates 100% similarity between the error and accurate clip and a score of 0% indicates zero similarities between the two caption clips. A greater similarity score would indicate greater caption quality. These scores are shown in the graph below. In the case of scoring differences between error and accurate captions, ACE seems to do better at mirroring participant perceptions. Overall, ACE is able to identify errors, but data from Q1 indicates that error identification does not necessarily reflect participant understanding. ACE and WER were especially inaccurate for clip number 5, as shown in Figure 16. Both metrics scored the clip high, above 80%, while participants found there to be a relatively large difference between the error and accurate captions, giving the clip a much lower score, less than 50%. Future research should work to identify the causes of these major discrepancies between metrics and human perception.



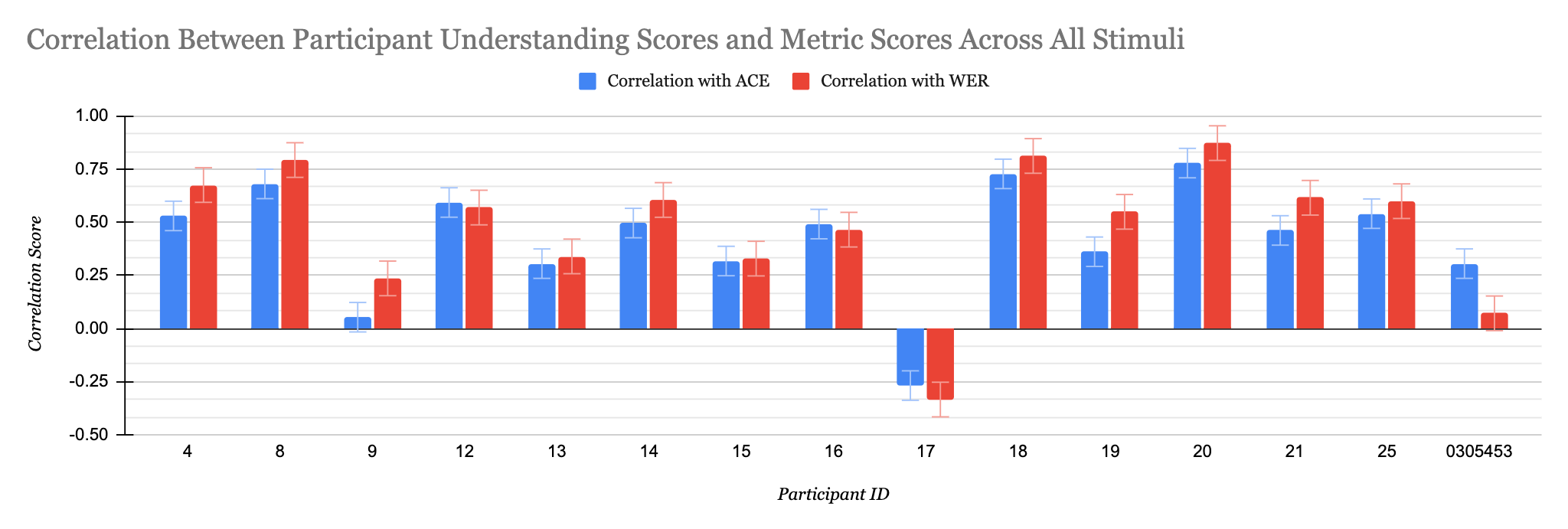
**Figure 15.** Above graphed is the ACE and WER metric score for each clip, along with the participants’ average understanding score.



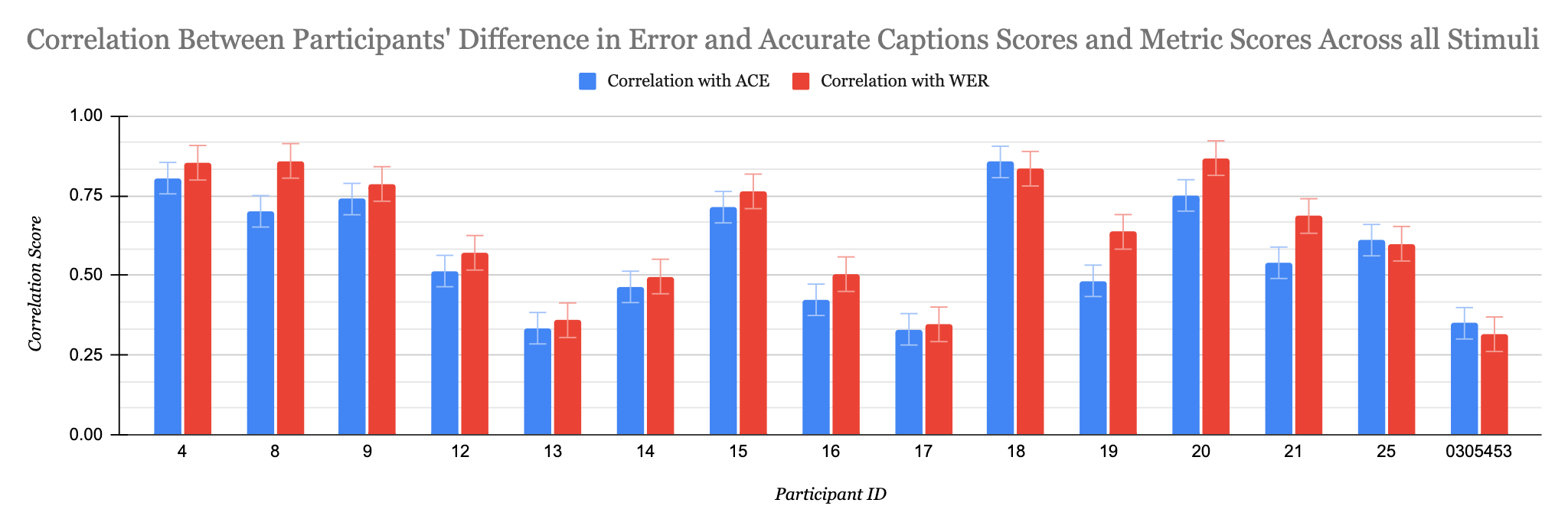
**Figure 16.** Above graphed is the ACE and WER metric score for each clip, along with the participants’ average difference between error and accurate captions score.

As aforementioned, participants on average reported that they were able to fully understand the original clip at a score of 5.37 out of 7. In this case, the ACE metric did not perform as well with respect to Q1, because participants generally did understand the stimuli, and ACE scored all stimuli relatively low. WER did not penalize errors as greatly and generated higher scores that seem to better reflect actual participant experiences. This relationship is shown in Figures 11-14, where the left side of both ACE scatterplots (Figures 11 and 13) shows extremely low correspondence and high variability. Lower ACE scores have contributed to less correlation between the ACE metric and participant responses in both Q1 and Q2. Overall, Figures 11-14 illustrate that participants responded to Q1 and Q2 with quite high variability, making it difficult to identify clear trends with either metric.

We performed a correlation analysis on the metric scores and Q1 responses for each participant, as well as metric scores and Q2 responses for each participant. Results can be viewed in Figure 17. WER had an average correlation of 0.48 with Q1, while ACE had an average of 0.43. With Q2, WER had an average correlation of 0.63, while ACE had an average of 0.57. WER did marginally better. As seen in Figure 17, participant 17’s responses have a negative correlation with both ACE and WER scores. Reasons for this individual’s divergence from the rest of the data could be that the individual understood almost all of the stimuli, regardless of captioning mistakes. Ratings for each clip ranged from 5 to 7 out of 7. Thus, correlation results from this participant did not follow a particular trend and tended toward the negative side. Many other participants had quite low correlations regarding Q1. On the other hand, Q2 seemed to have more success in creating consistent responses from participants. Still, both questions represent important aspects of quality in captioning, and were important conditions to test metrics against.



**Figure 17.** The correlations between each metric and participants’ understanding scores for Q1, analyzed across all 10 stimuli responses for each participant.



**Figure 18.** The correlations between each metric and participants’ scores of difference between error and accurate captions for Q2, analyzed across all 10 stimuli responses for each participant.

Two T Tests were performed on the data in Figures 17 and 18. One test compared the WER and ACE correlation data for Q1, and one compared the WER and ACE correlation data for Q2. The Q1 test had a value of 0.007, indicating a significant difference between how each metric correlated with participant data. The Q2 test had a value of 0.0030, and there was no significance between the two data sets. Although the T Test for the understanding scores showed significance between the WER and ACE data, the means of each correlation data set are still very similar. Thus, we cannot reject the null hypothesis.

**5. CONCLUSIONS AND FUTURE WORK**

Through a user study with fifteen participants, we compared two caption metrics systems, WER and ACE, for their accuracy in evaluating caption quality in live television. We compared human-perceived quality statistics with each caption metric’s data. Analysis of the correlation between human statistics and each caption metric found that WER had a slightly higher correlation with participants. Live television may be a less suitable environment for ACE than the types of stimuli used by prior work. We discovered some flaws in the ACE metric that have not been mentioned in any previous studies. The ACE metric tended to score clips with lower accuracy than what was perceived by participants, which limited its ability to correlate strongly with participants’ scores. However, the difference in performance between WER and ACE was not statistically significant. Data was highly varied, leading us to conclude that neither WER nor ACE are optimized for use with live television captioning.

While we were able to discover much about the performance of ACE and WER in an under researched setting, there were few limitations associated with our study. The link between participants’ self-reported scores and caption quality is not a direct one. Capturing accurate data from participants is difficult in the case of Q1 because rating understanding can have various factors and variations that do not necessarily reflect the caption quality. For instance, understanding scores may have been a reflection of content of the stimuli clips rather than the captions themselves. Future work could reframe the question in a way that would eliminate these factors and create more objective data. We were also restricted by time, and were not able to include WWER in our study.

In the future, our goal is to expand this work to other metrics such as WWER and NER. Future work should also explore how caption metrics could be better optimized for use with live television. More broadly, there is also work to be done to better define caption quality in metrics research. None of the study participants involved reported having a perfect experience with captions in the past. With new analysis of caption metrics, we hope to inform future revisions of ACE and WER so that caption metrics can better assess and improve captioning for all that use it.

**6. ACKNOWLEDGEMENTS**

The contents of this paper were developed in part under a grant from the National Science Foundation, grant #1757836 (REU AICT) and under a grant from the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR grant number #90DPCP0002). NIDILRR is a Center within the Administration for Community Living (ACL), Department of Health and Human Services (HHS). The contents of this poster do not necessarily represent the policy of NIDILRR, ACL, HHS, and you should not assume endorsement by the Federal Government. We thank Akhter Al-Amin and Mariana Arroyo Chavez for their help with calculating the caption metrics for the videos.

**7. REFERENCES**

[1] Jang, P. J., & Hauptmann, A. G. (1999, June). Improving acoustic models with captioned multimedia speech. In *Proceedings IEEE International Conference on Multimedia Computing and Systems* (Vol. 2, pp. 767-771). IEEE.

[2] Block, M. H., & Okrand, M. (1983). Real-time closed-captioned television as an educational tool. *American Annals of the Deaf*, *128*(5), 636-641.

[3] Apone, T., Brooks, M., & O’Connell, T. (2010). Caption Accuracy Metrics Project. *Caption Viewer Survey: Error Ranking of Real-time Captions in Live Television News Programs. Boston*.

[4] Al Amin, A. (2020). Audio-Visual Caption Evaluation Metric for People who are Deaf and Hard of Hearing.

[5] Apone, T., Botkin, B., Brooks, M. & Goldberg, L. (2011). Research into Automated Error Ranking of Real-time Captions in Live Television News Programs. *The Carl and Ruth Shapiro Family National Center for Accessible Media at WGBH (NCAM).*

[6] Kafle, S., & Huenerfauth, M. (2017, October). Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 165-174).

[7] https://code.google.com/archive/p/word2vec