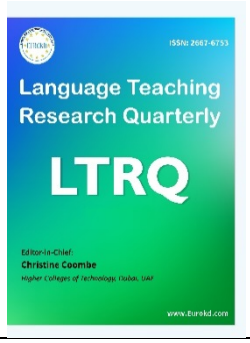



Language Teaching Research Quarterly

2022, Vol. 29, 65–91

# Using Statistical Transformation Methods to Explore Speech Perception Scale Lengths

Alyssa Kermad[1]*, Valeria Bogorevich[2]

[1]California State Polytechnic University, English and Modern Languages, Pomona, USA

[2]Arizona Western College, Modern Languages Division, Yuma, USA

***Received*** *22 August 2021* ***Accepted*** *16 April 2022*

**Abstract**

The practice of second language (L2) speech perception has traditionally relied on equal-interval perceptual scales and novice listeners' (NLs) impressionistic judgments of constructs such as accentedness and comprehensibility (Munro & Derwing, 2011). However, issues have surfaced with respect to how well NLs can use these scales, whether they use the entire scale, and how valid/reliable their ratings are. This study draws inspiration from the work of Glenn Fulcher in the area of L2 speech assessment, specifically on construct validity and scale design (1993; 2003). The current study explores similar issues from the lens of the speech perception practice, relying on scale transformation methods and multi-faceted Rasch measurement to examine the overall question of an ideal equal-interval scale length. Speech ratings of L2 accentedness and comprehensibility were originally collected from 56 NLs on 9-point bipolar Likert scales. Statistical transformations from 9-point to 7- and 5-point scales were compared through several indices (i.e., listener consistency, listener severity, scalar point discrimination, and scale usage) across four common speech tasks (i.e., read-aloud, spontaneous speech, elicited imitation, and picture description) (see Thomson & Derwing, 2015). With every statistical scale length reduction across four tasks, listeners' consistency improved, their levels of severity slightly decreased, and their scalar usage became more precise. Implications for L2 speech perception research suggest that the shorter 5-point perceptual scale shows potential in increasing the reliability and validity of scoring, while more appropriately distinguishing among speaker ability levels.

**Keywords:** *Speech Perception Scale Length, Speech Perception Scale Reliability, Pronunciation Scale*

adkermad@cpp.edu

doi:10.32038/ltrq.2022.29.05

The areas of second language (L2) speech perception and L2 speech assessment share similar practices in that they both seek to obtain scores which capture an L2 speaker's ability level on a given construct when operationalized on some sort of assessment artifact. At the same time, these two areas are also marked by distinguishable differences—the first being that L2 assessment relies on trained raters to obtain these scores, while L2 speech perception relies on novice listeners (NLs). The second major difference between the two fields is the very artifact of the listener/rater's judgments. L2 speech assessment typically makes use of criterion-oriented scales or analytic rubrics which carefully detail the skills and abilities that should be present at a particular band on the scale. Raters are intensively trained to use these scales, often relying on benchmark samples, calibration, recalibration, and discussion of incongruencies. On the other hand, in speech perception, the use of bi-polar Likert scales, especially the use of the 9-point scale, have become quite common to capture listeners' impressionistic judgments. Due to the limited dialogue in the past between L2 speech assessment and L2 speech perception researchers, in addition to the limited research on scalar operationalization and construct validity in speech perception, criterion referenced scales in speech perception have not commonly been adopted (Isaacs & Thomson, 2013). Perceptual scales seem to be appropriate when capturing a listener's global, impressionistic judgment, yet construct validity should still be taken into consideration.

Scale length likely plays a role in impacting the construct validity of important speech judgments. While Isaacs and Thomson (2013) critically compared 9-point and 5-point perceptual scales for comprehensibility, accentedness, and fluency, the ideal scale length remains unresolved as NLs have still been prone to undesirable listener variance due to their lack of training (Kang et al., 2019; Kermad, 2021). The current study attempts to increase the dialogue between speech perception and speech assessment researchers through the investigation of an "ideal" speech perception scale. The current study draws on assessment literature and analyses (i.e., multi-faceted Rasch measurement) to evaluate commonly used scalar points (9 vs. 7 vs. 5) (Isbell, 2017) across the most commonly used pronunciation tasks (see Thomson & Derwing, 2015). Drawing on speech assessment research, specifically on the topics of construct validity, rating behavior, and scalar functioning, these practices can inform speech perception researchers of the ideal scale length which can most optimally capture NLs' judgments of accentedness and comprehensibility.

## Review of the Literature

### *Speech Perception and Construct Validity*

Fulcher (1993, 1996a, 1996b, 1999) argued that the consideration of construct validity (see Messick, 1989) is vitally important in the process of speaking assessment. In addition, Fulcher (1995) and Fulcher and Márquez Reiter (2003) emphasized the importance of tests reflecting real-world situations. Later in 2007, Fulcher and Davidson prioritized validity among other important variables in assessment. From the argument-based approach to validity (Fulcher & Davidson, 2007; Kane 1992; Kane, 2013; Chapelle, 2012), a research instrument cannot be deemed valid unless there is evidence showing that the evaluation or scoring of a language performance sample is functioning well.

In order to capture valid ratings of accentedness and comprehensibility, it would be fruitful to ensure that the most optimal scale is being used and functioning as well as possible, especially when being used by novice listeners. In speech perception research, listeners' rating data have commonly been used as the outcome variable for L2 speech performance (Munro & Derwing, 2011), and their judgments have traditionally been captured on equal interval scales which allow for global impressions when a construct is placed on a metathetic continuum (Southwood & Flege, 1999). Accentedness (degree of accent) and comprehensibility (ease of understanding) have both been validated as constructs amenable to these scales (Munro, 2017; Southwood & Flege, 1999). The benefits of using NLs to judge L2 speech relates to their impressionistic, global judgments which can be reliable and sensitive to inherent properties in the stream of speech (Anderson-Hsieh & Koehler, 1988; Derwing et al., 2004; Isaacs & Thomson, 2013; Saito et al., 2016). Further, NLs can provide judgments which provide important insight into how well English learners are understood by members of a given L2 speech community (Munro, 2008). For these reasons, researchers in speech perception often recruit NLs with no linguistic training. At the same time, these types of Likert scales lack performance-based descriptions which then rely on NLs to create their own internal standards and boundaries of how scalar points align with ability levels. These standards, however, may vary from one listener to another.

*Rating Behavior*

In their discussion of possible threats to validity, Crooks et al. (1996) outlined how validity can be threatened when raters show undue emphasis on specific rubric criteria or unfairly favor particular response forms or styles. Furthermore, scores can be undermined when there is insufficient intra-rater/inter-rater consistency. Standardized testing companies (e.g., TOEFL, IELTS, etc.) undergo strict rater training to ensure that rater reliability is met. Indeed, speech assessment literature has modeled training procedures after standardized testing companies to evaluate the effect of training. Kang et al. (2019) found that TOEFL-modeled training drastically reduced listener variance and calibrated severe and lenient raters to similar levels of severity. Other speech assessment researchers (Kermad, 2021; Wei, 2015; Xi & Mollaun, 2011) also found the effects of rater training to be successful, especially when compared to groups without specialized training.

Formal rater training in speech perception research is not commonly practiced as it is in assessment. When listeners are trained in speech perception, it's often indicative of their background or experience, such as being graduate students in applied linguistics, teachers, native speakers of English, etc. (see Bongaerts et al., 1997; Crowther et al., 2015; Saito et al., 2016; Rossiter, 2009). However, despite the lack of formal training, it is common to obtain high reliability coefficients either via Cronbach's alpha or intra-class correlation coefficient (Isaacs & Thomson, 2013; Isbell, 2017; Kermad, 2021). These reliability coefficients take into consideration how the listeners function as a group, whereas other analyses (such as multi-faceted Rasch measurement (MFRM) reveal how raters perform individually along the lines of other rating indices. Indeed, when using these more fine-grained approaches, challenges with using these perceptual Likert scales have been uncovered (Isaacs & Thomson, 2013; Isbell, 2017; Kermad, 2021).

*Scale Functioning*

Speech assessment typically relies on analytic or holistic rubrics where each score has detailed or brief descriptions for each scoring band. Assessment researchers have been analyzing scales and rater performance from different perspectives. Fulcher (1996b) used discriminant analyses and Rasch partial credit analysis to assess a fluency scale, noting important considerations related to how the scale defines ability levels and how it represents actual language performance. Others have used FACETS' graphic output of the category probability curves in order to see how the scale scores function (see Myford & Wolfe, 2000; 2003; 2004; Li et al., 2019; Wind, 2020). One of the main concerns for a rating scale is whether each category on the scale comes to a peak. When categories do not peak, they have less probability of being assigned and should perhaps be collapsed with other categories (Myford, 2006). Based on such recommendations, TOEFL reduced the number of its rubric bands from seven to four (Jamieson & Poonpon, 2013). In addition, Fox and Jones (1998) noted that several issues can arise when raters have a large number of Likert points to choose from, such as clustering their answers in the middle and not using the endpoints, or not being able to distinguish between the points on the scale in a consistent and systematic way.

Speech perception commonly relies on these Likert scales for capturing listeners' impressionistic judgments of a construct (although sliding scales have also been integrated into the current practice; see Crowther et al., 2015; Saito et al., 2016). However, some issues have been noted with respect to how these scales function. For example, while a ceiling effect has been found with 7-point scales (Southwood & Flege, 1999), listeners have difficulty with 9-point scales, especially with distinguishing the middle points (Isaacs & Thomson, 2013). At the same time, even though the 5-point scale has been reported to be "constraining," the points have also been used more precisely than with a 9-point scale (Isaacs & Thomson, 2013, p. 147). Further exacerbating the situation is the extent to which novice listeners can reliably use these scales. While Isbell (2017) found high reliability coefficients with 9-point accentedness and comprehensibility scales, using MFRM, listeners were found to demonstrate differences in their rating consistency, in how they used the scales and what points they preferred, and in their severity/leniency.

**Research Questions**

The current study attempts to determine the scale length (9-, 7-, or 5-point) which functions the most ideally with NLs and their ratings of accentedness and comprehensibility across the four most commonly used tasks in L2 pronunciation (see Thomson & Derwing, 2015). Statistical transformation methods using MFRM provide detailed indices regarding listener rating behavior and scale functioning. These statistical checkpoints are analyzed across the different scale lengths in order to decide on the most optimally functioning scale. The current study is therefore guided by the two following research questions: (1) *On which speech perception scale (9 vs. 7 vs. 5) do novice listeners demonstrate the most ideal rating behavior (i.e., consistency and severity) for accentedness and comprehensibility across four tasks?* And *(2) Which speech perception scale (9 vs. 7 vs. 5) yields the most ideal scale functioning (i.e., point discrimination and scalar usage) for accentedness and comprehensibility across four tasks?*

## Methods

*Participants*

Speakers. Speaker data was the source of speech stimuli to be presented to NLs. A total of fifteen Intensive English Program (IEP) students (seven males; eight females; age range 18-30 years) were recruited to provide speech data. These students came from three different levels (low intermediate, intermediate, upper intermediate) of an IEP program in the Southwestern United States. At the time of data collection, five students were enrolled in Level 3, four in Level 4, and six in Level 5. These levels were determined by the IEP's robust placement test, which is structured to mimic the TOEFL iBT proficiency test. The corresponding TOEFL iBT scores for each level (from 3 to 5) are as follows: 32-44, 45-56, and 57-69. Three language backgrounds (i.e., Chinese, Japanese, and Arabic) and four geographical regions were represented: China, ($n$ = 6), Japan ($n$ = 1), Kuwait ($n$ = 3), and Saudi Arabia ($n$ = 5).

Listeners. Fifty-six NLs (22 males; 34 females; age range 18-22 years) were recruited to rate the speech data for accentedness and comprehensibility. These were undergraduate students enrolled in required composition classes at a four-year university in the Southwestern United States. Most of these students were in their first semester at this university and represented diverse, inter-disciplinary majors (i.e., criminology, math, English, undeclared, etc.). Forty-seven of these listeners were English NSs and 9 were highly proficient NNSs. Screening tests (i.e., 12 Mann-Whitney U tests (for unequal $n$ sizes) to compare NS's and NNSs' ratings for each construct across four tasks) revealed no significant differences between the ratings of the NSs and the NNSs in this study.

*Speech Stimuli*

Listener ratings of accentedness and comprehensibility were obtained across the four most commonly used tasks in pronunciation research (Thomson & Derwing, 2015); read-aloud tasks, spontaneous speech tasks, elicitation tasks, and picture description tasks. The read-aloud task was a brief paragraph about earthquakes in California, taken from Celce-Murcia et al. (2010, p. 328). This task was appropriate for comprehensively eliciting a range of segmental and suprasegmental features. For the spontaneous speech task, speakers were given a prompt which asked them to speak for approximately one minute about differences in culture between their home city and the American city in which they currently lived. Celce-Murcia et al. (2010) suggest using spontaneous speech tasks to elicit speech which comes naturally to the speakers; in this case, the topic of cultural differences was one that yielded readily accessible and comfortable monologic speech. The elicited imitation (from Trofimovich et al., 2009) was a list of six sentences about a clown and his dog. This task was appropriate for providing a target model, as the researcher read the sentences aloud, and the speaker repeated and recorded the sentences immediately after the researcher. The final task was an 8-scene picture narrative (Derwing et al., 2009), commonly drawn upon in L2 pronunciation research, which depicted a man and a woman who accidentally exchanged suitcases after having bumped into each other on a busy street corner.

*Rating Scales*

All original listener data were collected on two 9-point scales for comprehensibility and accentedness following common practice (see Isbell, 2017; Munro & Derwing, 1995a; 1995b).

For comprehensibility, listeners used a semantic left-side positive Likert scale of 1 (extremely easy to understand) and 9 (impossible to understand). For accentedness, the scale was similar, except with different descriptions (i.e., 1 signified no strong accent and 9 signified a very strong accent). Each listener heard all 60 speech samples (15 speakers x 4 tasks) and provided their impressionistic ratings. This resulted in 3,360 overall scores for each construct (56 listeners x 4 tasks x 15 speakers). Listener data were not collected on the 7-point and 5-point perceptual scales, but instead, statistical methods were used to transform the original data from the 9-point scales to these shorter scales (see "Data Analyses" below).

*Procedures*

One-on-one meetings were held with the speaker participants who recorded their responses to all four tasks successively using the freeware program Praat (Boersma & Weenink, 2016) and a noise-cancelling microphone. To prepare the files for the listener judgments, any background noise audible at the beginning of the file was removed and replaced with one second of generated silence. In order to keep the content as similar as possible, the first 12 seconds were taken from each file for each task. Twelve seconds was long enough to hear an adequate sample of speech, yet short enough to make an immediate judgment. Twelve seconds was also longer than the average 7 seconds used in Derwing and Munro (1997). The 60 files were randomized and presented to the listeners through the online survey program, Survey Gizmo, (https://www.surveygizmo.com/). The listeners provided their comprehensibility and accentedness ratings after each file in the survey interface. No breaks were offered, and the entire listener survey lasted approximately 45 minutes.

*Data Analyses*

Multi-faceted Rasch measurement (MFRM) using a computer program FACETS, version 3.71.4 (Linacre, 2015) was used to examine how NLs performed when providing impressionistic ratings of comprehensibility and accentedness for speakers across tasks. The data were analyzed using the following variables, which are called facets: speaker ($n$ = 15), listener (novice listeners; $n$ = 56), task (four tasks; $n$ = 4), and criteria (accentedness and comprehensibility; $n$ = 2). FACETS allows for the investigation of patterns of listener characteristics by placing raw scores on a log odds scale of equal-interval units (McNamara, 1996).

Several iterations of scalar recoding were performed, beginning with the original data obtained from the 9-point scale. Following Linacre's (1999) guidelines, scalar points were combined by analyzing the base statistics of the 9-point scale and attempting to logically combine points of low scalar usage or unclear scalar points to statistically generate new 7- and 5-point scales. Most often this was done by combining two scalar points into one. Once the scale transformations were obtained, rating behavior and scale functioning were evaluated for each scale to determine the ideal rating scale length.

## Results

One of FACET's first points of inspection is its variable maps. In this case, the variable maps (see Figures 1-3 below) place the four facets (i.e., speaker, listener, task, and rating criteria) on the same logit scale. The variable maps can be interpreted in the following manner:

(1) The *first column* labelled "Measure" provides the logit scale in which average ability/difficulty/severity is set at 0 logits for all facets, excluding the listener facet which was non-centered and allowed to float. In other words, the severity estimates of listeners were relative to speakers and tasks.

(2) The *second column*, labelled "Speaker," represents the speaker facet. This facet is negative due to the nature of the scales in this study. The raters were instructed to award lower numbers to those who had more comprehensible and less accented speech. Thus, it was necessary to make the speaker facet negative in order to let the program know that the students who received lower numbers had higher ability levels and should be placed at the top of the column; whereas the students who were assigned higher numbers had lower ability levels and consequently needed to be positioned at the bottom.

(3) The *third column*, "Listeners," illustrates the severity/leniency of the 56 listeners. More severe listeners are at the top and more lenient ones are at the bottom.

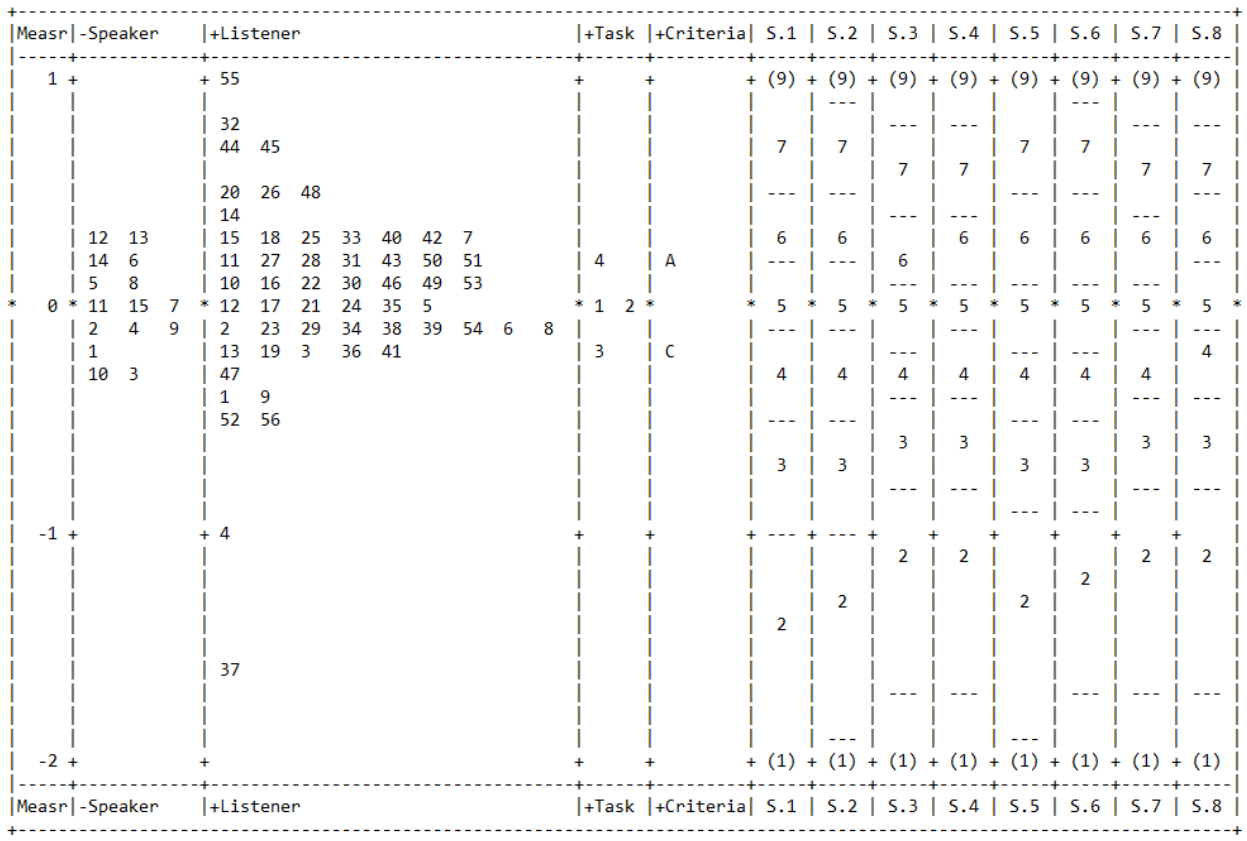
(4) The *fourth column*, "Task," is the task facet, where number 1 is the read-aloud task, 2 is the spontaneous speech task, 3 is the elicitation task, and 4 is the picture description task. Tasks above 0 are more difficult, and those below zero are less difficult.

(5) The *fifth column*, "Criteria," indicates the criteria facet with "A" representing accentedness and "C" comprehensibility. Constructs above zero were rated more severely, and constructs below zero were rated more leniently.

(6) The *last eight columns* represent how the scalar points were utilized for comprehensibility or accentedness across four tasks (see *Note* under Figures *1-3* for a description of abbreviations). The longer the depiction of a scale for a number, the more often it was used. It should be noted that due to the combination of scalar points on the 7- and 5-point scales, not all numbers on these scales are represented chronologically on the variable maps.

**Figure 1**

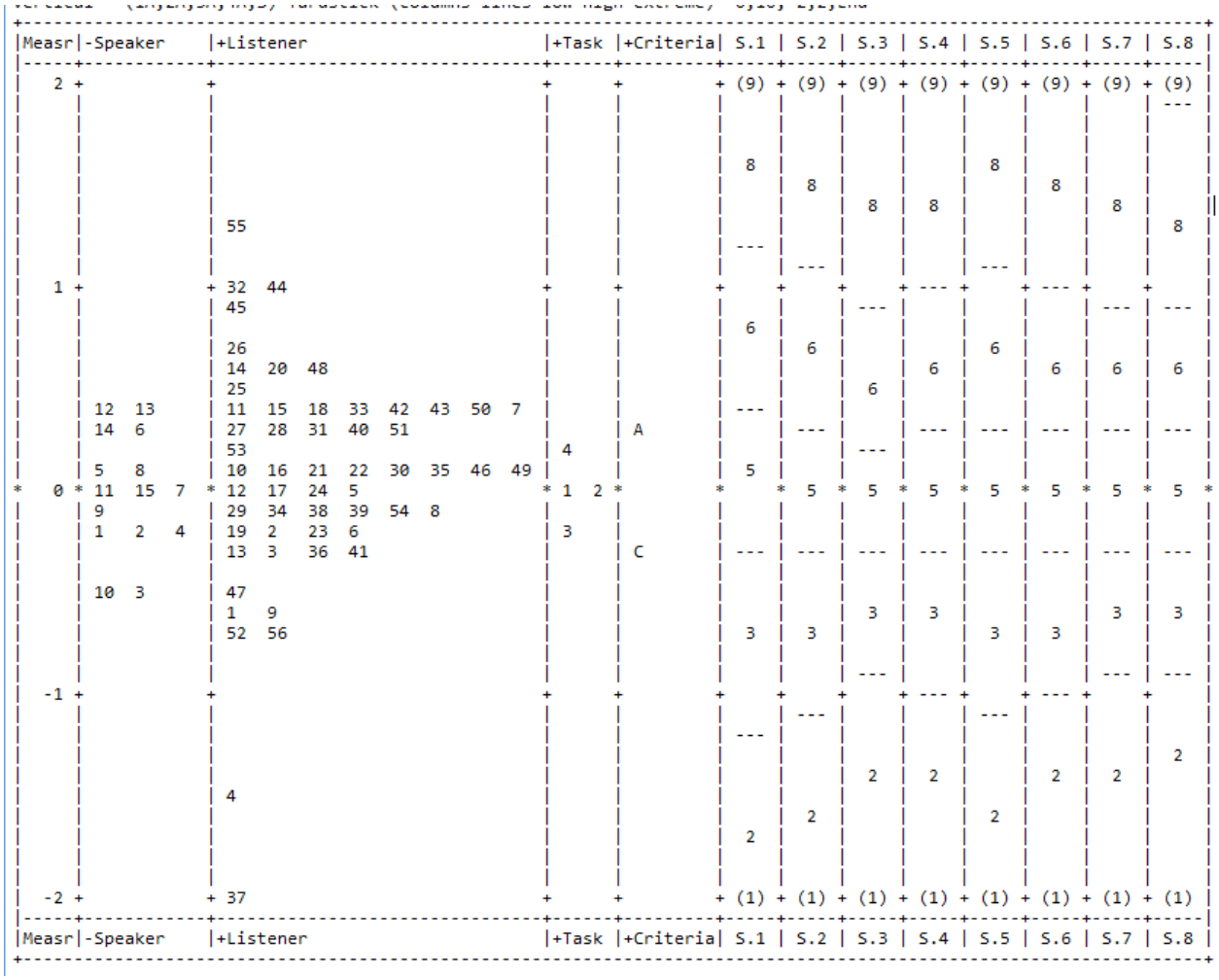
*Variable Map for 9-point Scale*




Note: S.1 = Comprehensibility, Task 1; S.2 = Accentedness, Task 1;
S.3 = Comprehensibility, Task 2; S.4 = Accentedness, Task 2;
S.5 = Comprehensibility, Task 3; S.6 = Accentedness, Task 3;
S.7 = Comprehensibility, Task 4; S.8 = Accentedness, Task 4.

**Figure 2**

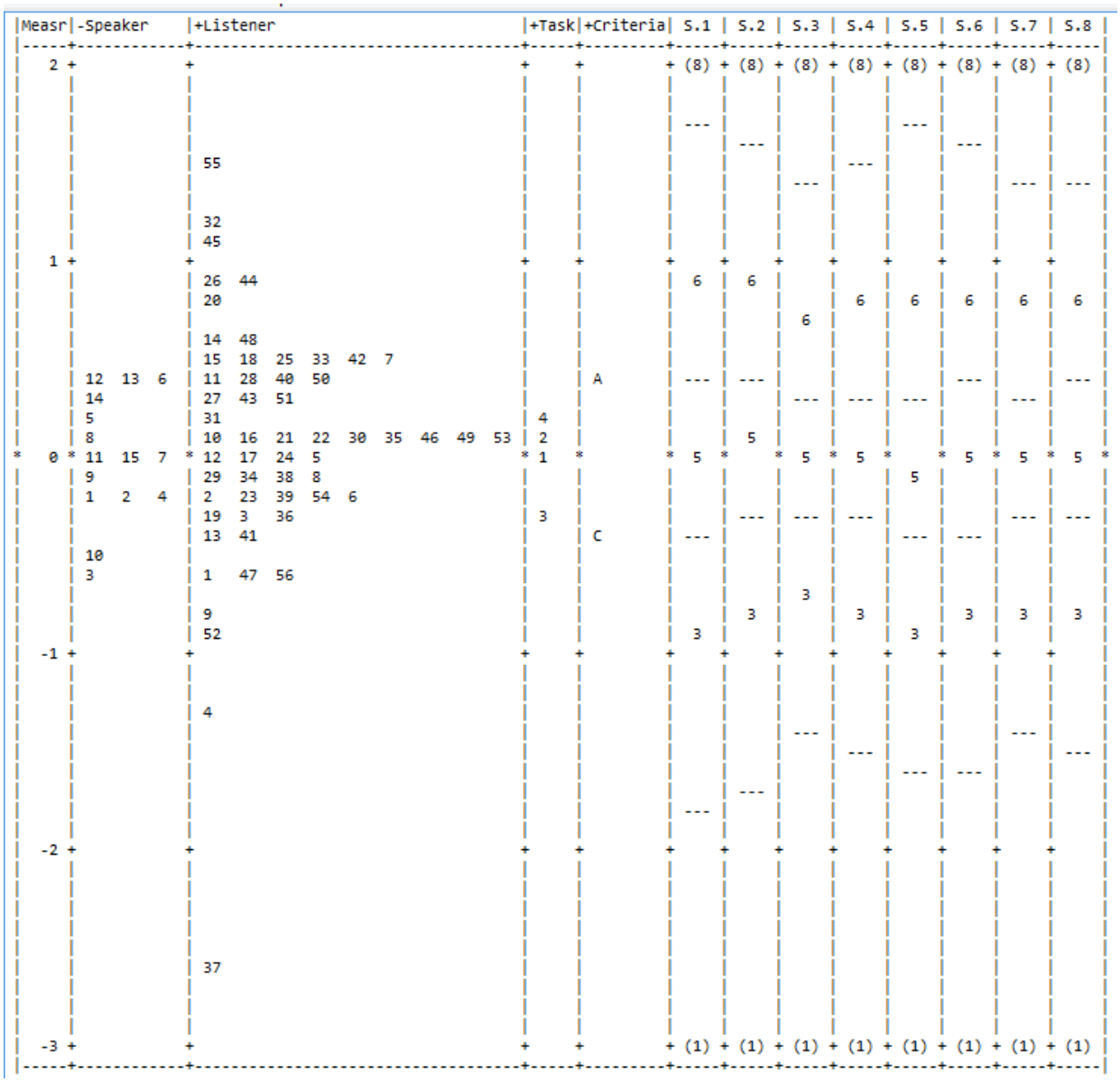
*Variable Map for 7-point Scale*




Note: S.1 = Comprehensibility, Task 1; S.2 = Accentedness, Task 1;
S.3 = Comprehensibility, Task 2; S.4 = Accentedness, Task 2;
S.5 = Comprehensibility, Task 3; S.6 = Accentedness, Task 3;
S.7 = Comprehensibility, Task 4; S.8 = Accentedness, Task 4.

**Figure 3**

*Variable Map for 5-point Scale*




Note: S.1 = Comprehensibility, Task 1; S.2 = Accentedness, Task 1;
S.3 = Comprehensibility, Task 2; S.4 = Accentedness, Task 2;
S.5 = Comprehensibility, Task 3; S.6 = Accentedness, Task 3;
S.7 = Comprehensibility, Task 4; S.8 = Accentedness, Task 4.

*Rating Behavior*

To investigate listener rating behavior (consistency and severity), two indices were used: fit statistics and severity separation levels. Listener consistency was evaluated based on fit statistics (infit and outfit mean squares). Fit statistics show to what degree each listener exhibits self-consistency or whether listeners award scores predictably or erratically (Linacre, 1999), diverging from the expected pattern predicted by the model (Weigle, 1998). A fit statistic of 1 is considered

to be ideal. Different limits have been set for acceptable ranges of fit statistics, but for a low-stakes context (such as for the current data), the limits of 1.5 and .5 were adopted (Lunz et al., 1990). Therefore, listeners with values higher than the limit of 1.5 showed misfit, tending to score speakers' performance in an erratic, unpredictable way. On the other hand, listeners with values lower than the limit of .5 showed overfit, tending to be overly consistent by assigning similar scores to speakers of different abilities. Following the analysis of the fit statistics, listener severity statistics were examined in conjunction with the FACETS variable maps. Separation indices illustrated the levels of severity within the entire group of listeners for each scale.

Rating behavior on the 9-point scale. Illustrated in Table 1 are the listener identification numbers, severity logits, error statistics, fit statistics (i.e., infit mean square values), and correlations for the 9-point scale. Listeners with acceptable fit statistics were removed from the table for purposes of conciseness. According to the fit statistics, eight listeners showed misfit with values over 1.5, and two listeners showed overfit with values below .5.

**Table 1**

*9-point Scale Measurement Report for Listeners (Arranged by Infit Values)*

| Listener | Severity logit | Model error | Infit mean square | Correlation |
|---|---|---|---|---|
| 4 | **-1.00** | **.07** | **3.66** | **-.26** |
| 44 | **.70** | **.06** | **2.20** | **.44** |
| 53 | **.11** | **.05** | **2.01** | **.73** |
| 11 | **.25** | **.05** | **1.96** | **.57** |
| 52 | **-.54** | **.06** | **1.70** | **-.02** |
| 43 | **.24** | **.05** | **1.61** | **.57** |
| 31 | **.15** | **.05** | **1.59** | **.80** |
| 13 | **-.23** | **.05** | **1.52** | **.06** |
| 51 | .21 | .05 | 1.39 | .32 |
| -- | -- | -- | -- | -- |
| 54 | -.12 | .05 | .53 | .72 |
| 28 | **.21** | **.05** | **.43** | **.56** |
| 21 | **.03** | **.05** | **.38** | **.53** |
| *M* | .05 | .05 | 1.03 | .48 |
| *SD* | .40 | .01 | .54 | .20 |

Note: Reliability = .98; Separation: 7.30; Fixed chi-square: 2063.2 ($df$ = 55; $p$ < .001); -- indicate listeners with acceptable fit statistics that were removed from the table for purposes of conciseness

The separation index of 7.30 with reliability of .98 illustrates that the listeners reliably exercised slightly over seven different levels of severity, which was confirmed by significant fixed chi-square statistic ($\chi 2$ = 2063.2, $df$ = 55, $p$ < .001). These results suggest that the listeners were not interchangeable and rated the same speakers with different levels of severity for both comprehensibility and accentedness.

Rating behavior on the 7-point scale. The 9-point scale underwent several trial transformations to find the best fitting 7-point scale. This was done by recoding different combinations of scalar points on the original 9-point scale to find the most ideal 7-point scale based on the recommendations in Linacre (1999). It was clear that some mid-points on the 9-point scale were not clearly discriminating ability levels. Therefore, for the best version of the 7-point scale, the points 3 and 4 were combined, in addition to the points 6 and 7.

Table 2 displays the listener fit statistics for the final 7-point sale. Using the same cut-off values as with the 9-point scale, six listeners showed misfit by exceeding the value of 1.5, and three listeners showed overfit with values below .5. While still not ideal, these fit statistics were a slight improvement when compared with the original 9-point scale.

**Table 2**

*7-point Scale Measurement Report for Listeners (Arranged by Infit Values)*

| Raters | Severity logit | Model error | Infit mean square | Correlation |
|---|---|---|---|---|
| **4** | **-1.50** | **.09** | **3.34** | **-.25** |
| **53** | **.17** | **.07** | **2.37** | **.71** |
| **44** | **1.05** | **.08** | **2.23** | **.45** |
| **11** | **.35** | **.07** | **2.03** | **.59** |
| **43** | **.37** | **.07** | **1.68** | **.58** |
| **31** | **.26** | **.08** | **1.68** | **.80** |
| 56 | -.66 | .08 | 1.50 | .55 |
| -- | -- | -- | -- | -- |
| 20 | .60 | .08 | .50 | .42 |
| **15** | **.39** | **.07** | **.49** | **.54** |
| **28** | **.30** | **.07** | **.42** | **.53** |
| **21** | **.08** | **.07** | **.42** | **.50** |
| *M* | .07 | .08 | 1.02 | .47 |
| *SD* | .55 | .01 | .55 | .19 |

Note: Reliability = .98; Separation: 7.11; Fixed chi-square: 2214.8 ($df$ = 55; $p$ < .001); -- indicate listeners with acceptable fit statistics that were removed from the table for purposes of conciseness

The separation index of 7.11 with reliability of .98 illustrates that the listeners reliably exercised over seven levels of severity, which was confirmed by a significant fixed chi-square statistic ($\chi 2$ = 2214.8, $df$ = 55, $p$ < .001). Similar to the 9-point scale, these results suggest that the listeners were not interchangeable and rated the same speakers significantly differently for both comprehensibility and accentedness. This index was only slightly reduced on the 7-point scale.

Rating behavior on the 5-point scale. Finally, to investigate the effectiveness of the 5-point scale, several iterations of recoding (following recommendations in Linacre, 1999) were performed in order to find the most ideal scale. In the end, the following combinations of the original 9-point scale maximized reliability, combining points 1-2, 3-4, 6-7, and 8-9. Listener fit statistics are presented in Table 3. Five listeners demonstrated misfit with infit values above 1.5,

and one listener demonstrated overfit with a fit value below .5. These fit statistics once again demonstrated a small improvement from both the original 9-point and transformed 7-point scales.

**Table 3**

*5-point Scale Measurement Report for Listeners (Arranged by Infit Values)*

| Raters | Severity logit | Model error | Infit mean square | Correlation |
|---|---|---|---|---|
| **4** | **-1.30** | **.11** | **2.86** | **-.23** |
| **52** | **-.90** | **.10** | **2.06** | **-.05** |
| **44** | **.94** | **.10** | **1.70** | **.42** |
| **11** | **.38** | **.08** | **1.63** | **.54** |
| **13** | **-.40** | **.08** | **1.55** | **.07** |
| 43 | .35 | .08 | 1.46 | .55 |
| -- | -- | -- | -- | -- |
| 15 | .48 | .09 | .60 | .52 |
| 28 | .38 | .08 | .51 | .52 |
| **21** | **.11** | **.08** | **.45** | **.50** |
| *M* | .07 | .09 | 1.02 | .46 |
| *SD* | .62 | .02 | .40 | .19 |

Note: Reliability = .98; Separation: 6.76; Fixed chi-square: 1701.6 (df = 55; p < .001); -- indicate listeners with acceptable fit statistics that were removed from the table for purposes of conciseness

Once again, the statistics in conjunction with the variable map displayed in Figure 7 were used to examine listener severity for the 5-point scale. The separation index was less than both the 9-point and 7-point scales (6.76) with reliability of .98, illustrating that the listeners reliably exercised less than seven levels of severity. This was confirmed by significant fixed chi-square statistic ($\chi 2 = 1701.6$, $df = 55$, $p < .001$).

*Scale Functioning*

Scale functioning took into consideration both how the actual points on the scale discriminated speakers' ability levels and how much of the scale was actually being used. Figures 4 to 6 present the probability curves that depict the structure of the scoring scale. Probability curves are a visual representation of the probability of a certain score being given to a speaker based on that speaker's ability level (Wright & Masters, 1982; Linacre, 1999). The horizontal axis depicts the two constructs, accentedness and comprehensibility, in equal interval logits; the vertical axis shows the probability of receiving a score on the rating scales for a speaker at a given ability level. Because of the nature of the rating scale, negative logits on the x-axis depict higher abilities of accentedness and comprehensibility. Numbers which do not come to distinct peaks illustrate no clear probability of a speaker receiving a specific score for their level of comprehensibility or accentedness. Ideally, each point on the rating scale should come to a distinct peak, showing a clear alignment between scores and ability levels and therefore ideal scale functioning.

Twenty-four probability curves were obtained and inspected, one for each construct ($n = 2$), for each task ($n = 4$), for three scales. Figures 4-6 present the probability curves only for the read-

aloud task (the most common task in the pronunciation literature; see Thomson & Derwing, 2015) as a point of illustration. The complete set of probability curves can be found in Appendix A. Overall trends of the probability curves will be discussed in the following sections for all tasks.

Scale functioning on the 9-point scale. Figure 4 displays the probability curves for the read-aloud task for comprehensibility and accentedness. Probability curves for an ideally functioning scale have score points that peak successively; however, this is not the case for the 9-point scale. It can be seen that the numerical points are blended together with no separately discernible peaks for a scalar point. This indicates that hardly any category is clearly most probable for the speakers' ability levels. The points 1 and 9 are extrapolated by the model. The situation is largely the same for both comprehensibility and accentedness for all four tasks (see Appendix A), demonstrating that listeners struggled to identify particular ability levels using the 9-point scale, regardless of the task. These patterns suggest that listeners assigned similar scores to speakers of different accentedness and comprehensibility ability levels. For example, a speaker of mid-proficiency comprehensibility (0 logits) had the probability of being awarded multiple comprehensibility scores ranging from 1 to 9.

**Figure 4**

*Comprehensibility (Left) and Accentedness (Right) Ratings of 9-point Scale on Read-aloud Task*

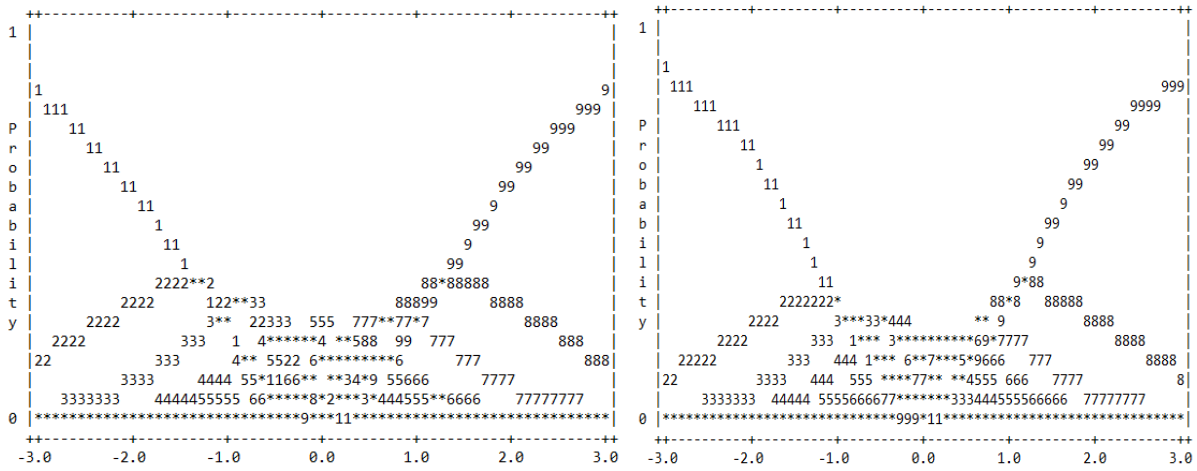



Table 4 illustrates the percentages of how often each scalar point was selected by listeners for all tasks. For all tasks, listeners assigned the endpoints (1 and 9) minimally. For the rest of the scalar points, no point was used more than 20% of the time.

**Table 4**

*Percentages of Scale Usage for 9-point Scale*

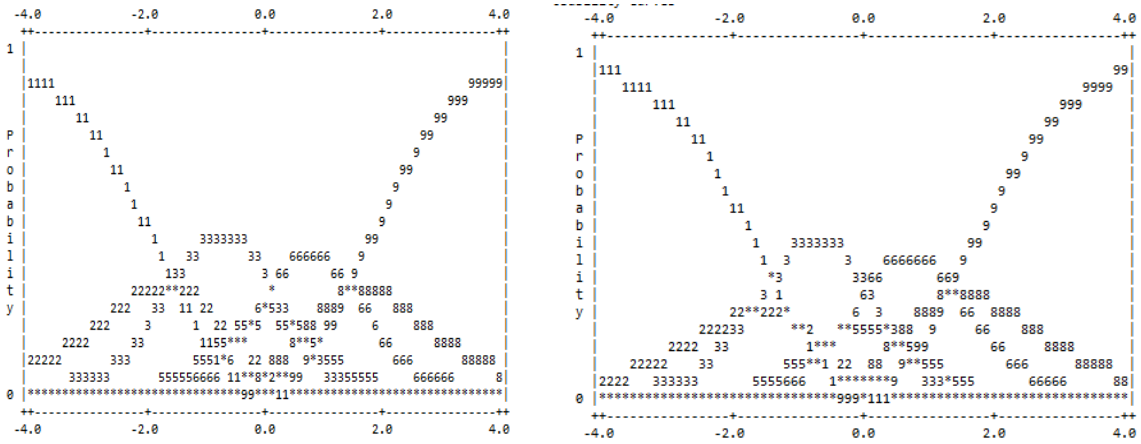
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| | Comprehensibility | | | | | | | | |
| Read-aloud | 6% | 12% | 18% | 17% | 18% | 11% | 10% | 6% | 2% |
| Spontaneous | 10% | 13% | 14% | 14% | 14% | 12% | 11% | 9% | 4% |
| Elicitation | 11% | 17% | 18% | 16% | 15% | 11% | 8% | 3% | 1% |
| Picture | 6% | 10% | 12% | 13% | 15% | 13% | 14% | 10% | 7% |
| | Accentedness | | | | | | | | |
| Read-aloud | 3% | 4% | 8% | 13% | 16% | 17% | 16% | 14% | 9% |
| Spontaneous | 3% | 4% | 7% | 9% | 14% | 18% | 17% | 15% | 14% |
| Elicitation | 4% | 6% | 9% | 14% | 15% | 20% | 14% | 10% | 8% |
| Picture | 3% | 4% | 7% | 9% | 14% | 15% | 16% | 14% | 17% |

Note: Comprehensibility:1= extremely easy to understand; 9 = extremely difficult to understand; Accentedness: 1 = no strong accent; 9 = a very strong accent

Scale functioning on the 7-point scale. Figure 5 displays the probability curves for the read-aloud task for comprehensibility and accentedness on the transformed 7-point scale. Because the points 3-4 and 6-7 were combined, they became represented by 3 and 6 respectively. Comparing the probability curves on the 7-point scale to those of the 9-point scale, some improvements can be noted. The peaks of 3 and 6 begin to come to more pronounced curves. Nevertheless, the numeric points are still not evenly spaced out and are crowded underneath each other. For example, the scalar point 6 begins to peak but consumes other points below it, illustrating that several points were being assigned to speakers at a given ability level. Visual inspections of the probability curves across tasks (see Appendix A) illustrate similar patterns for the four tasks, yet with slightly cleaner mid points for the spontaneous and elicitation tasks.

**Figure 5**

*Comprehensibility (Left) and Accentedness (Right) Ratings of 7-point Scale on Read-aloud Task*




The percentages of scale usage for the 7-point scale are displayed in Table 5. For all tasks, the new point of 3 on the 7-point scale became the most frequently used for comprehensibility, and

the new point of 5 became the most frequently used for accentedness. This suggested that listeners were still hovering within the inner-bounds of the scales.

**Table 5**

*Percentages of Scale Usage for 7-point Scale*

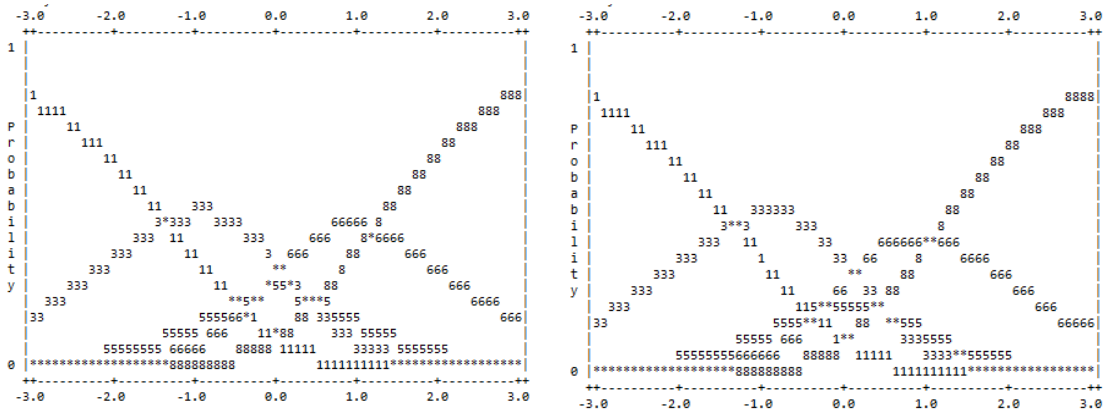
| **Transformed scale** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
|---|---|---|---|---|---|---|---|
| Combined points | 1 | 2 | 3 (+4) | 5 | 6 (+7) | 8 | 9 |
| | Comprehensibility | | | | | | |
| Read-aloud | 6% | 12% | 36% | 18% | 21% | 6% | 2% |
| Spontaneous | 10% | 13% | 28% | 14% | 13% | 9% | 4% |
| Elicitation | 11% | 17% | 35% | 15% | 18% | 3% | 1% |
| Picture | 6% | 10% | 25% | 15% | 27% | 10% | 7% |
| | Accentedness | | | | | | |
| Read-aloud | 3% | 4% | 22% | 16% | 33% | 14% | 9% |
| Spontaneous | 3% | 4% | 16% | 14% | 35% | 15% | 14% |
| Elicitation | 4% | 6% | 23% | 15% | 34% | 10% | 8% |
| Picture | 3% | 4% | 16% | 14% | 31% | 14% | 17% |

Note: Comprehensibility:1 = extremely easy to understand; 9 = extremely difficult to understand; Accentedness:1 = no strong accent; 9 = a very strong accent

Scale functioning on the 5-point scale. Figure 6 displays the probability curves for the read-aloud task for comprehensibility and accentedness on the transformed 5-point scale. The 5-point scale was created by leaving point 5 as is and combining points 1-2, 3-4, 6-7, and 8-9, represented by 1, 3, 6, and 8 respectively. The probability curves on the 5-point scale are a large improvement from the 9- and 7-point scales. The curves come to more distinct peaks and are more spaced out. The point of 5, however, does not peak and is submerged underneath most of the other points, likely due to its mid-position. Similar to the 9-point and 7-point scales, visual inspections of the probability curves for the four tasks (see Appendix A) did not reveal any pronounced difference in point discrimination across tasks.

**Figure 6**

*Comprehensibility (Left) and Accentedness (Right) Ratings of 5-point Scale on Read-aloud Task*




The percentages of scale usage for both accentedness and comprehensibility are displayed in Table 6 for the transformed 5-point scale. The scalar point of 1 had an increased usage — more so than with the 9- and 7- points scales. The middle scores of the five-point scale also demonstrate increased usage. The highest points on the scale were used more for accentedness, showing that listeners rated accent more severely than comprehensibility.

**Table 6**

*Percentages of Scale Usage for 5-point Scale*

| **Transformed scale** | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|
| Combined points | 1(+2) | 3(+4) | 5 | 6(+7) | 8(+9) |
| | Comprehensibility | | | | |
| Read-aloud | 18% | 36% | 18% | 21% | 8% |
| Spontaneous | 22% | 28% | 14% | 23% | 13% |
| Elicitation | 28% | 35% | 15% | 18% | 4% |
| Picture | 15% | 25% | 15% | 27% | 18% |
| | Accentedness | | | | |
| Read-aloud | 7% | 22% | 16% | 33% | 23% |
| Spontaneous | 7% | 16% | 14% | 35% | 29% |
| Elicitation | 10% | 23% | 15% | 34% | 18% |
| Picture | 7% | 16% | 14% | 31% | 31% |

Note: Comprehensibility: 1= extremely easy to understand; 9= extremely difficult to understand; Accentedness: 1= no strong accent; 9= a very strong accent

## Discussion

This study used statistical transformation methods to examine rating indices on three commonly used scales (9-point, 7-point, and 5-point) in speech perception research. NLs provided impressionistic ratings of non-native speech across the four most commonly used tasks in the pronunciation literature (see Thomson & Derwing, 2015). For all transformed scales, the same data check points for rating behavior and scale functioning were examined, including listener fit and severity statistics, probability curves, and scale usage. For rating behavior, the transformed 5-point scale yielded the most consistent ratings with slightly fewer levels of severity when

compared to the 7-point and 9-point scales. In fact, there were only six listeners who showed misfit/overfit on the 5-point scale; this was compared to nine listeners on the 7-point scale and ten listeners on the 9-point scale. The 5-point scale was created by combining points on the scale which were not clearly working to differentiate between ability levels or those which were not clearly represented on the 9-point scale. With fewer scalar points, results suggest that listeners have the potential to establish a more concrete one-to-one correspondence of scalar points with speakers' ability levels.

Both the transformed 7-point and 5-point scales only slightly decreased the different levels of severity with which scores were assigned. That is, combining points from the 9-point scale did not seem to be adequate enough in calibrating listeners to similar severity levels. Likely, the length of the scale is simply not enough to do this. In Kermad (2021), it took a 2-hour multi-step training procedure to improve raters' judgments of accentedness, comprehensibility, and rated listener comprehension. Similarly, Kang et al. (2019) found that it took a stringent rating session modeled after TOEFL training procedures to calibrate a subset of raters of extreme severity and extreme leniency to similar levels of severity. A host of listener background variables can come into play when making impressionistic judgments of accented speech, including accent familiarity (see Fayer & Krasinksi, 1987; Harding, 2012; Major et al., 2002; Scales et al., 2006), topic familiarity (see Gass & Varonis, 1984), stereotyping (Kang & Rubin, 2009), or native speaker status (Kang, 2012; Kang et al., 2019). In absence of training, these factors can play a role in the severity/leniency with which listeners assign scores.

The scale which worked the best to discriminate among the proficiency levels of the speakers, ultimately leading to more valid scores (Fulcher, 1993, 1996a, 1996b, 1999; Linacre, 1999), was also the transformed 5-point scale. Compared to the 9-point scale, the differences in scale functioning were fairly dramatic. On the 9-point scale, for example, a speaker with a mid-level ability for both comprehensibility and accentedness had the chance of being awarded multiple scores on the scale ranging from 1 to 9. This suggests that listeners lacked agreement on how the scale should be used and which points should be assigned to speakers of different ability levels. This echoes Isaacs and Thomson's (2013) findings that raters had a hard time distinguishing between so many numbers on the 9-point scale. Results of this study were consistent with Isbell's findings (2017) as well in that scalar points were subsumed, overlapping, and non-defining. At the same time, while the 5-point scale was a stark improvement from both the 9- and 7-point scales, some middle points were still muddled, suggesting once again that some sort of rating intervention, such as benchmark samples, may still be needed for ideal scalar functioning. A further suggestion to improve the listener-scale interaction is applying more descriptors to the points on the scale, such as in Kermad (2021). Perhaps, bi-polar descriptions are insufficient for listeners to parse the mid-points of the scale. Drawing upon Fulcher's work in speech assessment (Fulcher, 2003; Fulcher, 2015), speech perception can integrate the best practices of speech assessment in linking construct definition with rating scale development.

For scale usage, the 9-point and 7-point scales illustrated endpoints which were used minimally by the listeners, a similar trend found in Isaacs and Thomson (2013). Yet when combining points on the 5-point scale, more scores of 1 and 5 became allocated for both accentedness and

comprehensibility. Furthermore, the 7- and 5-point scales yielded higher concentrations of scores at certain mid points; however, the 9-point scale did not demonstrate these patterns. While it is important to keep in mind that these discussions are drawn from statistical transformation methods, it is noteworthy in demonstrating the potential of all points on the shorter scale to be used more frequently.

While the four most common pronunciation tasks were used in this study to obtain a more comprehensive picture about how these scales functioned, listeners did not seem to perform any "better" or "worse" in terms of the task. In other words, the patterns of listener consistency, severity, scalar discrimination, and scale usage were largely replicated across tasks. This illustrates that although speakers may perform differently across tasks, listeners may be affected similarly by the length of the scale regardless of the task.

## Conclusion

While research continues to evolve with respect to how listeners use scales in their perception of non-native speech, the current study has attempted to preliminarily address the issue of the "ideal" scale length for gathering NLs' impressionistic judgments of L2 speech. Knowing that the ideal scale is never really "ideal," these findings have provided initial evidence that the shorter 5-point scale may function most effectively by reducing listener inconsistency and slightly decreasing levels of severity, while at the same time improving the correspondence between scores and ability levels and providing a more balanced distribution of scalar usage. At the same time, these implications should be considered cautiously alongside the limitation that data for this study were only gathered on the 9-point scale, and it was only through statistical transformation methods that these rating indices were provided. Therefore, until further research can validate the results through data collection on all three scales, the current study only reveals the potential of using shorter rating scales for impressionistic judgments. Rubric design in speech perception may be a part of the future (see Kermad, 2021), and the measurement-driven evidence discussed in this study can eventually be backed up by performance data (Fulcher, 1987; 1993; 1996b; 2003; Fulcher et al., 2011), i.e., phonological characteristics of accentedness and comprehensibility to differentiate speakers' given ability levels at set scores.

## References

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning*, *38*, 561-613. https://doi.org/10.1111/j.1467-1770.1988.tb00167.x

Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer (Version 6.0.11) [Software]. Available from http://www.fon.hum.uva.nl/paul/praat.html

Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition*, *19*(4), 447-465. https://doi.org/10.1017/S0272263197004026

Celce-Murcia, M., Brinton, D. M., Goodwin, J. M., & Griner, B. (2010). *Teaching pronunciation: A course book and reference guide*. Cambridge University Press.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…, *Language Testing, 29*, 19-27. https://doi.org/10.1177/0265532211417211

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessment. *Assessment in Education*, *3*(3), 265-285. https://doi.org/10.1080/0969594960030302

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *The Modern Language Journal*, *99*(1), 80-95. https://doi.org/10.1111/modl.12185

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*(1), 1-16. http://www.jstor.org/stable/44488664

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, *31*(4), 533-557. https://doi.org/10.1017/S0272263109990015

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*(4), 655-679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*(3), 313-326. https://doi.org/10.1111/j.1467-1770.1987.tb00573.x

Fox, C. M., & Jones, J. A. (1998). Use of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology, 45*(1), 30-45. https://doi.org/10.1037/0022-0167.45.1.30

Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a foreign language* [Unpublished Ph.D. Dissertation]. University of Lancaster, UK.

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. English *Language Teaching Journal, 41*(4), 287-291. https://doi.org/10.1093/elt/41.4.287

Fulcher, G. (1995). Variable competence in second language acquisition: a problem for research methodology? *System, 23*(1), 25-33. https://doi.org/10.1016/0346-251X(94)00055-B

Fulcher, G. (1996a). Testing tasks: Issues in task design and the group oral. *Language Testing, 13*(1), 23-51. https://doi.org/10.1177/026553229601300103

Fulcher, G. (1996b). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing,13*, 208-238. https://doi.org/10.1177/026553229601300205

Fulcher, G. (1999). Assessment in English for academic purposes: putting content validity in its place. *Applied Linguistics, 20*, 221-36. https://doi.org/10.1093/APPLIN/20.2.221

Fulcher, G. (2003). *Testing second language speaking*. Longman/Pearson Education.

Fulcher, G. (2015). Assessing second language speaking. *Language Teaching, 48*(2), 198-216. https://doi.org/10.1017/S0261444814000391

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment*. Routledge.

Fulcher, G., & Marquez Reiter, R. (2003). Task difficulty in speaking tests. *Language Testing, 20*(3), 321-344. https://doi.org/10.1191/0265532203lt259oa

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5-29. https://doi.org/10.1177/0265532209359514

Gass, S., & Varonis, F. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34*(1), 56-89. https://doi.org/10.1111/j.1467-1770.1984.tb00996.x

Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing, 29*(2), 163-180. https://doi.org/10.1177/0265532211421161

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*(2), 135-159. https://doi.org/10.1080/15434303.2013.769545

Isbell, D. (2017). Assessing pronunciation for research purposes with listener-based numerical scales. In O. Kang & A. Ginther (Eds.), *Assessment in Second Language Pronunciation* (pp. 89-111). New York: Routledge.

Jamieson, J., & Poonpon, K. (2013). Developing analytic rating guides for TOEFL IBT's integrated speaking tasks. *ETS Research Report Series, 2013*(1), i-93. https://doi.org/10.1002/j.2333-8504.2013.tb02320.x

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly, 9(*3), 249-269. https://doi.org/10.1080/15434303.2011.642631

Kang, O., & Rubin, D. L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology, 28*(4), 441-456. https://doi.org/10.1177/0261927X09341950

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing, 36*(4), 481-504. https://doi.org/10.1177/0265532219849522

Kermad, A. (online 2021). Training the "everyday" listener how to rate accented speech. *International Journal of Listening*. https://doi.org/10.1080/10904018.2021.1987910

Li, S., Taguchi, N., & Xiao, F. (2019) Variations in rating scale functioning in assessing speech act production in L2 Chinese.*Language Assessment Quarterly, 16*(3), 271-293. https://doi.org/10.1080/15434303.2019.1648473

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103-122.

Linacre, J. M. (2015). *Facets* (Version No. 3.71.4). MESA Press. https://www.winsteps.com/facgood.htm

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331-345. https://doi.org/10.1207/s15324818ame0304_3

Major, R., Fitzmaurice, S., Bunta, F., & Balasubramanian. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly, 36*(2), 173-190. https://doi.org/10.2307/3588329

McNamara, T. F. (1996). *Measuring Second Language Performance*. Longman. https://doi.org/10.2307/330236

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-104). American Council on Education and Macmillan.

Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 193-218). John Benjamins Publishing Company. https://doi.org/10.1075/sibil.36

Munro, M. J. (2017). Dimensions of pronunciation. In Kang, O., Thomson, R. I. & Murphy, J. (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 413-431). Routledge. https://doi.org/10.4324/9781315145006

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73-97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*, 289- 306. https://doi.org/10.1177/002383099503800305

Munro, M. J., & Derwing, T. M. (2011). The foundations of accent and intelligibility in pronunciation research. *Language Teaching, 44*(3), 316-327. https://doi.org/10.1017/S0261444811000103

Myford, C. M., & Wolfe, E. W. (2000). Monitoring sources of variability within the Test of Spoken English assessment system (TOEFL Research Report No. 65). Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2000.tb01829.x

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement, 4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Myford, C. (2006). *Analyzing rating data using Linacre's Facets computer program: A set of training materials to learn to run the program and interpret output* [Unpublished manuscript].

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review, 65*(3), 395-412. https://doi.org/10.3138/cmlr.65.3.395

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, *37*, 217-240. https://doi.org/10.1017/S0142716414000502

Scales, J., Wennerstrom, A., Richard, D., & Wu, S. H. (2006). Language learners' perceptions of accent. *TESOL Quarterly*, *40*(4), 715-738. https://doi.org/10.2307/40264305

Southwood, M. H., & Flege, J. (1999). Scaling foreign accent: Direct magnitude estimation versus interval scaling. *Clinical Linguistics & Phonetics, 13,* 335-349. https://doi.org/10.1080/026992099299013

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics, 36*(3), 326-344. https://doi.org/10.1093/applin/amu076

Trofimovich, P., Lightbown, P. M., Halter, R. H., & Song, H. (2009). Comprehension-based practice. *Studies in Second Language Acquisition, 31*(4), 609-639. https://doi.org/10.1017/S0272263109990040

Wei, J. (2015). *Assessing speakers of world Englishes: The roles of rater language background, language attitude and training* [Unpublished doctoral dissertation].

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287. https://doi.org/10.1177/026553229801500205

Wind, S. A. (2020). Do raters use rating scale categories consistently across analytic rubric domains in writing assessment? *Assessing Writing*, *43*, 100416. https://doi.org/10.1016/j.asw.2019.100416

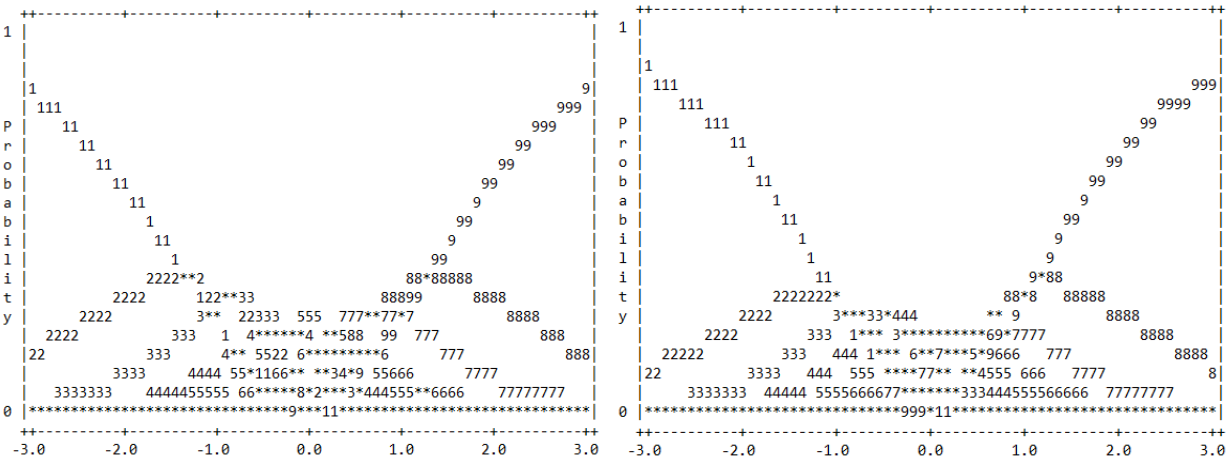
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement.* MESA Press.

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning, 61*(4), 1222-1255. https://doi.org/10.1111/j.1467-9922.2011.00667.x

## Appendix A

### Figure 7

*Comprehensibility (Left) and Accentedness (Right) Ratings of 9-point Scale on Read-aloud Task*




### Figure 8

*Comprehensibility (Left) and Accentedness (Right) Ratings of 9-point Scale on Spontaneous Speech Task*

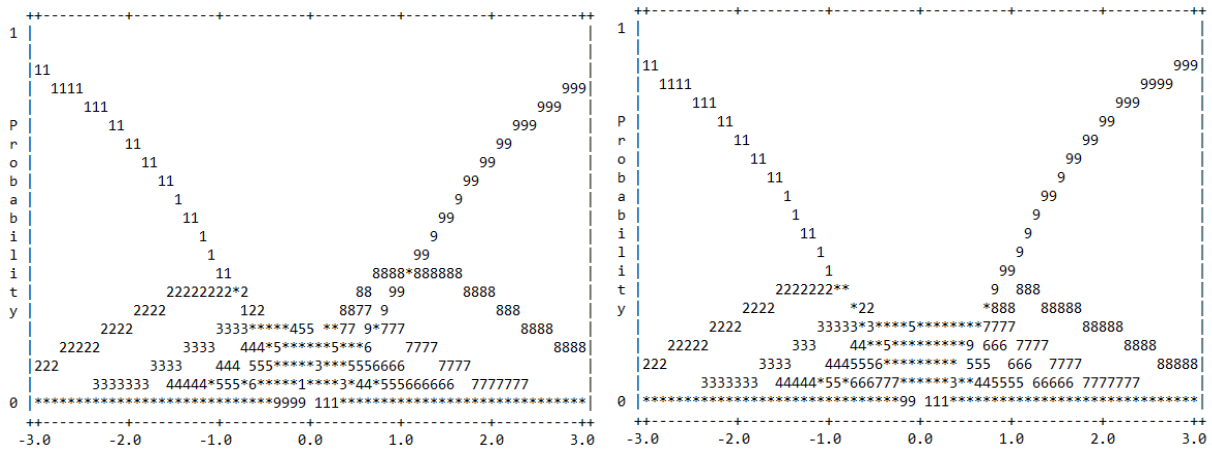


**Figure 9**

*Comprehensibility (Left) and Accentedness (Right) Ratings of 9-point Scale on Elicitation Task*

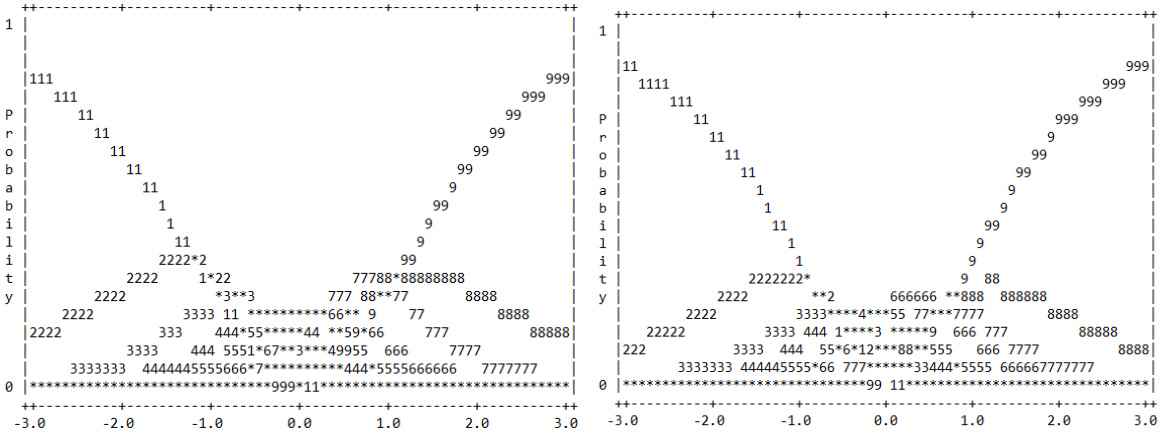


**Figure 10**

*Comprehensibility (Left) and Accentedness (Right) Ratings of 9-point Scale on Picture Description Task*

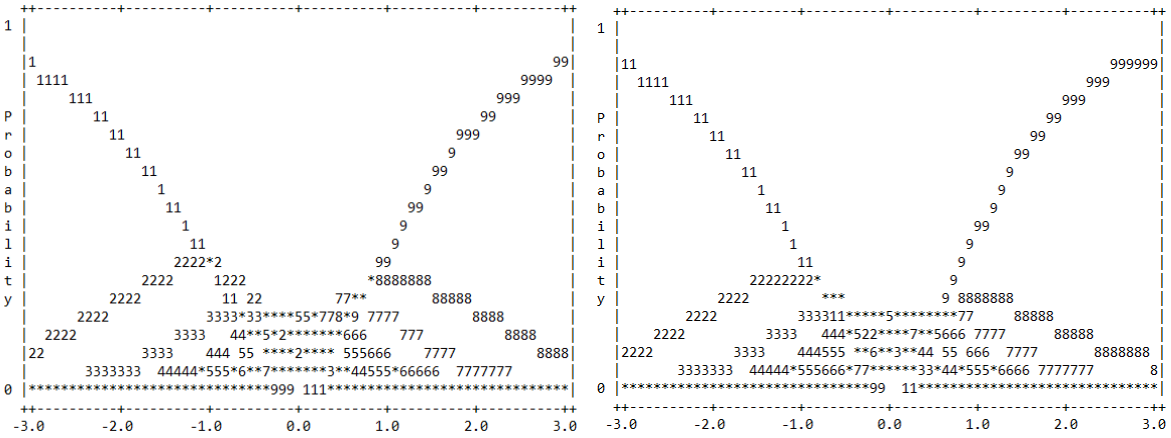


**Figure 11**

*Comprehensibility (Left) and Accentedness (Right) Ratings of 7-point Scale on Read-aloud Task*

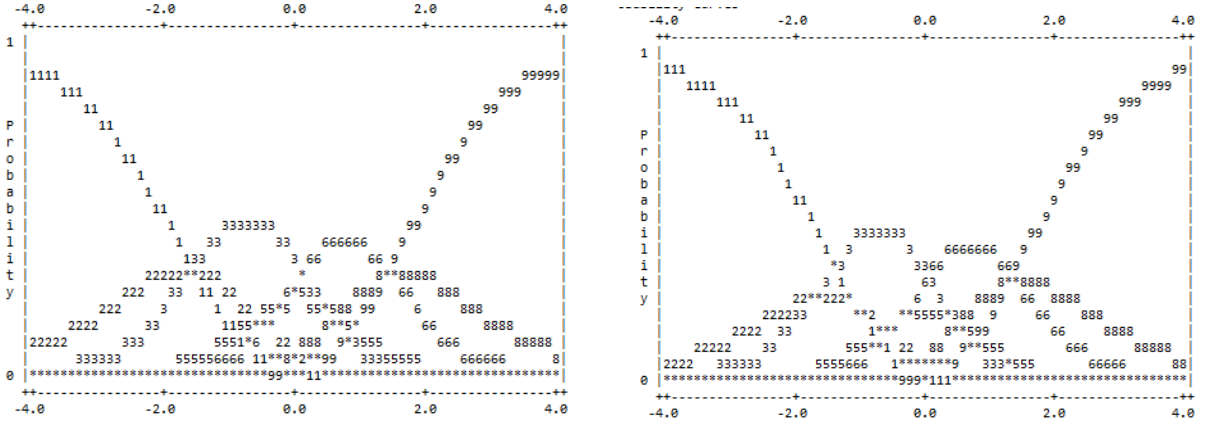



**Figure 12**

*Comprehensibility (Left) and Accentedness (Right) Ratings of 7-point Scale on Spontaneous Speech Task*

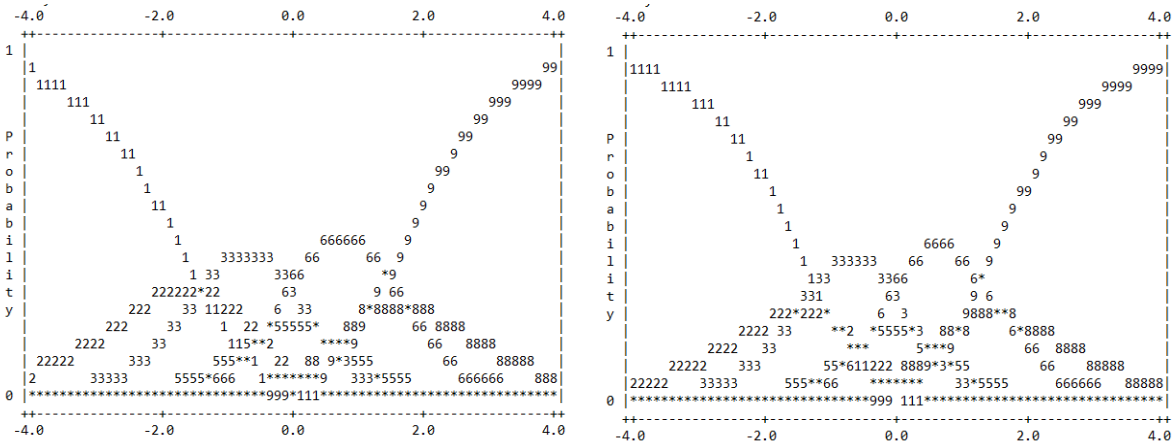



**Figure 13**

*Comprehensibility (Left) and Accentedness (Right) Ratings of 7-point Scale on Elicitation Task*

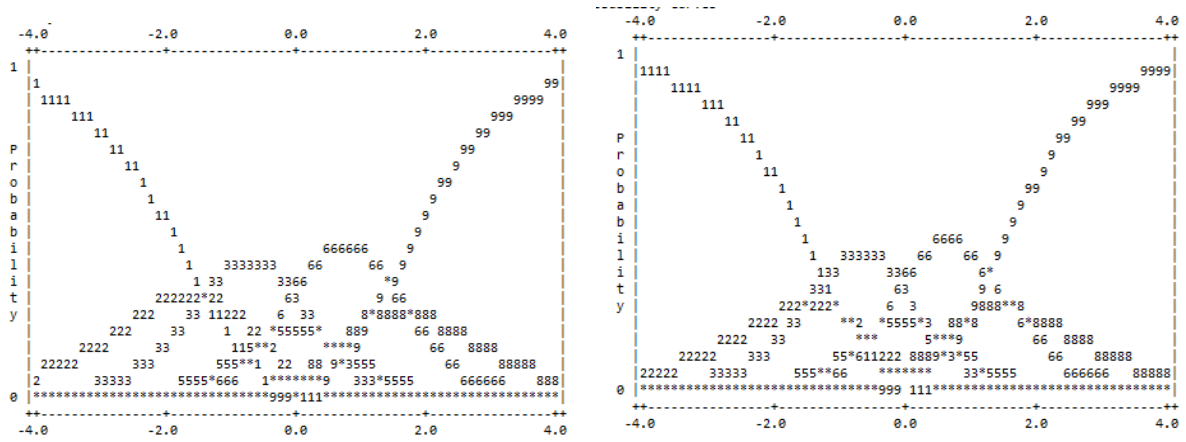


**Figure 14**

*Comprehensibility* (*Left*) and *Accentedness* (*Right*) *Ratings* of 7-point *Scale* on *Picture Description Task*

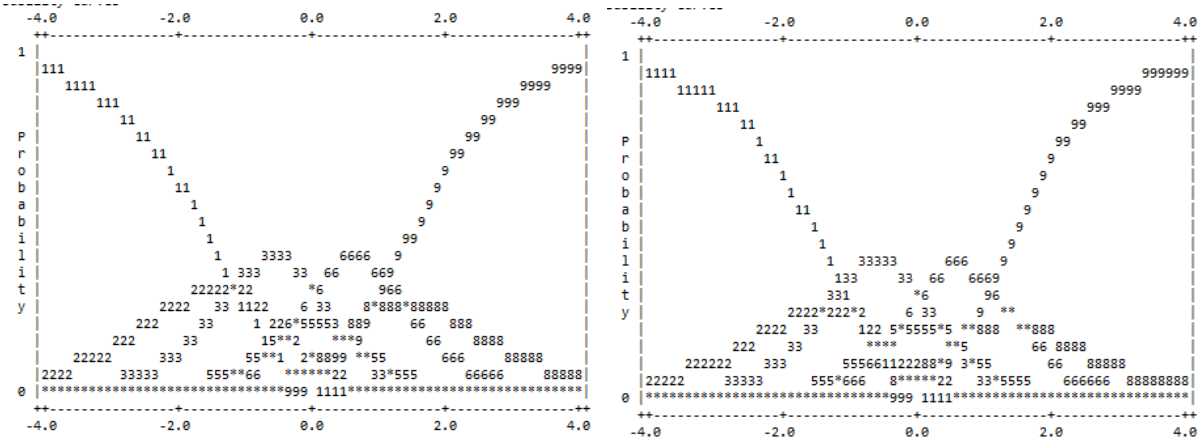



**Figure 15**

*Comprehensibility (Left) and Accentedness (Right) Ratings of 5-point Scale on Read-aloud Task*

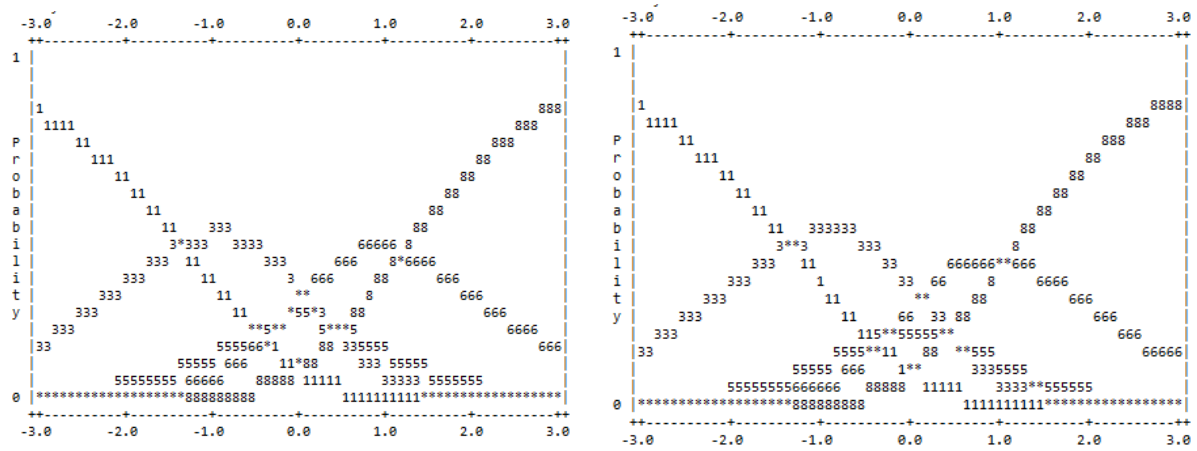


**Figure 16**

*Comprehensibility (Left) and Accentedness (Right) Ratings of 5-point Scale on Spontaneous Speech Task*

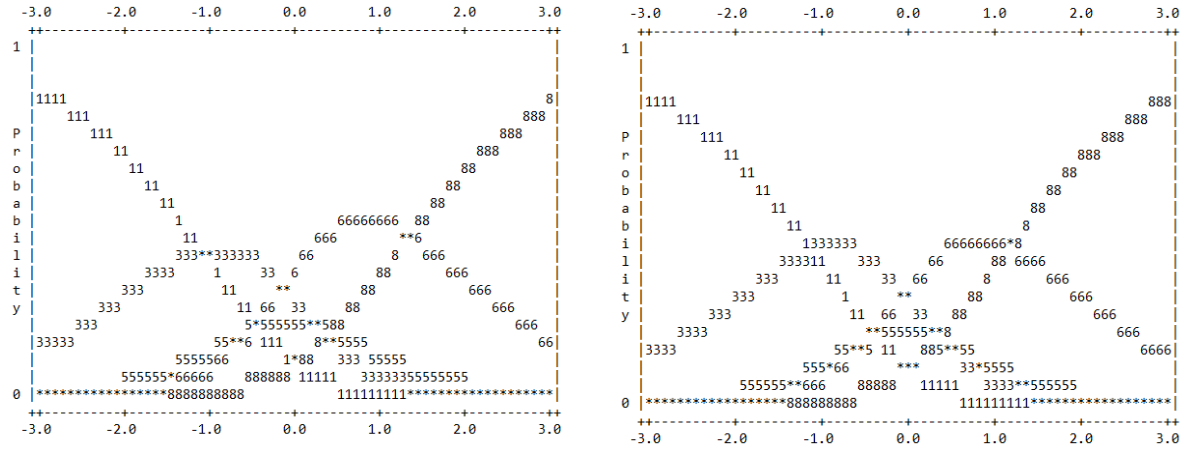



**Figure 17**

*Comprehensibility* (*Left*) and *Accentedness* (*Right*) *Ratings* of 5-point *Scale* on *Elicitation Task*

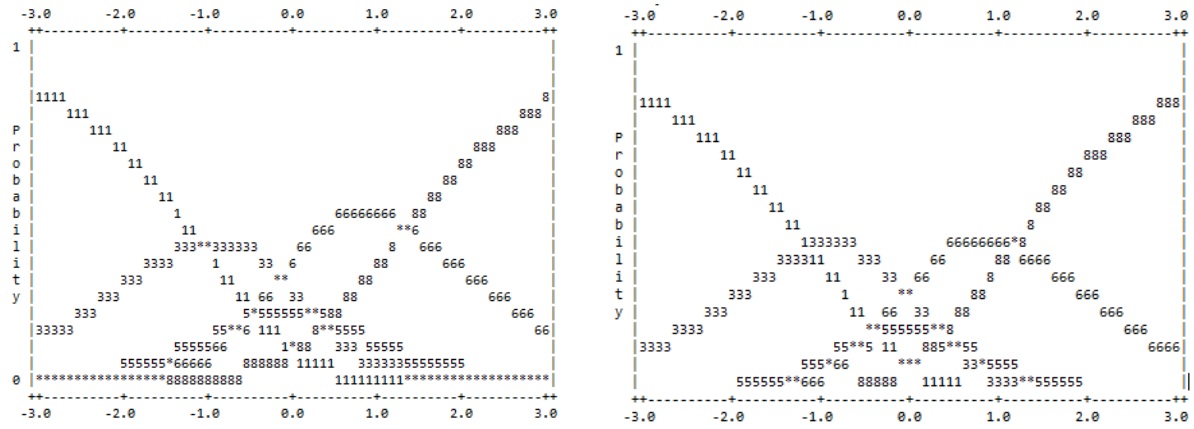


**Figure 18**

*Comprehensibility (Left) and Accentedness (Right) Ratings of 5-point Scale on Picture Description Task*


**Acknowledgements**

Not applicable.

**Funding**

Not applicable.

**Ethics Declarations**

**Competing Interests**

No, there are no conflicting interests.

**Rights and Permissions**

**Open Access**

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. You may view a copy of Creative Commons Attribution 4.0 International License here: http://creativecommons.org/licenses/by/4.0/.