Scaling foreign accent: direct magnitude estimation versus interval scaling

M. HELEN SOUTHWOOD and
JAMES E. FLEGE

University of Alabama at Birmingham, AL, USA

(*Received 5 November 1997; accepted 28 August 1998*)

Abstract

Scaling of degree of perceived foreign accent is common in studies of second language acquisition. Although interval scales are commonly used, it is not known whether accentedness is amenable to linear partitioning (i.e. described by a metathetic continuum) or whether accentedness is resistant to linear partitioning because it has degrees of magnitude or quantity (i.e. described by a prothetic continuum). This study determined if accentedness was a prothetic or metathetic continuum. Two groups of 10 native English listeners rated, in a self-paced experiment, degrees of perceived accent of native Italian speakers of English using both a seven-point equal-appearing interval scale and direct magnitude estimation (DME). DME scores were plotted against interval scores. This study found that accentedness (at least for Italian speakers) is amenable to linear partitioning. However, a wide dispersion of DME scores for '7' on the interval scale suggests a ceiling effect. A nine- or 11-point scale may better measure degree of perceived foreign accent. Individual listeners could reliably judge accentedness; however, inter-judge reliability was poor, possibly due to differences in listeners' internal standards of foreign accent or scaling artifact. Response biases occurred because foreign accent lacks a defined physical referent.

Keywords: foreign accent, scaling, direct magnitude estimation, interval scales.

Introduction

Foreign accent in second language speakers has received considerable attention (Flege, Munro and MacKay, 1995; Flege and Fletcher, 1992; Purcell and Suter, 1980). Munro (1998) defined foreign accent as 'non-pathological speech produced by second language learners that differs in partially systematic ways from the speech characteristics of native speakers of a given dialect'. The detection of foreign accent relates to acoustic differences between native and non-native speakers' segmental

Address correspondence to: M. Helen Southwood, PhD, University of Alabama at Birmingham, Division of Speech and Hearing Sciences, VH 503, Birmingham, AL 35293, USA. e-mail: mhsouthw@uab.edu

Clinical Linguistics & Phonetics
ISSN 0269-9206 print/ISSN 1464-507X online ©1999 Taylor & Francis Ltd
http://www.tandf.co.uk/JNLS/clp.htm
http://www.taylorandfrancis.com/JNLS/clp.htm

articulations, suprasegmental, and subsegmental levels (Flege, 1984). Voice onset times differ between non-native and native speakers of English (Flege, 1984; Williams, 1980; Flege and Hillenbrand, 1984). Differences in vowel durations have also been reported between native and non-native speakers of English (Mack, 1982). Other work has shown that foreign accent can be detected in sentences from which segmental identity has been removed digitally, but in which variations in rhythm and intonation have been preserved (Flege *et al.*, 1995). However, there is uncertainty as to how these speaker differences translate into listeners' perceptions of foreign accent. Identifying speaker differences is a perceptual event, and the listeners' reactions validate any measurements used to determine such differences (Young, 1969).

The perception of foreign accent leads to the realization that the talker is not a fellow native speaker. Native speakers can often detect foreign accent in the speech of a non-native talker after hearing a few syllables. Foreign accent can be detected in syllables, individual vowels and consonants, or in just 30-millisecond excerpts of word-initial consonants (Flege, 1984). Detection of foreign accent by listeners has several consequences. A foreign accent may affect the perceptions of the non-native talker as a person. Non-native talkers receive less positive ratings on dimensions related to social status (e.g. estimated professional success, education, intelligence, and wealth) (Brennan and Brennan, 1981). Listeners might give lower ratings to foreign accented speech because it is difficult to understand (Sebastian, Ryan, Keogh and Schmidt, 1980). In addition, difficulty processing the speech of the non-native talker might increase the cognitive load of the listener.

Global ratings of foreign accent assist in making predictions about the particular variables that influence foreign accent. Studies have shown that age of exposure to the second language influences the authenticity of the non-native talker's pronunciation (Flege *et al.*, 1995). Further, global ratings of accent permit identification of the particular acoustic variables that influence listeners' judgements of degree of perceived foreign accent. Perceptual measures are important because they corroborate the influence of these particular acoustic variables on listeners' perceptions of degree of perceived foreign accent and verify listeners' reactions to degree of foreign accent. To obtain constructive measures of degree of perceived foreign accent it is imperative that appropriate perceptual techniques be used.

Scaling techniques allow us to obtain global measures of foreign accent and measure the degree to which the accents of non-native speakers diverge from that of native speakers of a language. A standard scale has yet to be developed for such ratings and a variety of methods are currently in use (Munro, 1993). Degree of perceived foreign accent is commonly scaled using equal-appearing interval scales. These range from four-point scales to seven- or nine-point scales (Asher and Garcia, 1969; Oyama, 1982). Direct magnitude estimation (DME) has been used occasionally to scale foreign accent (Brennan, Ryan and Dawson, 1975; Ryan, Carranza and Moffie, 1977). Continuous scales have also been used (Flege, 1988; Munro, 1993).

Although a variety of scaling techniques have been used to judge foreign accent, there is uncertainty as to which scale is the most appropriate. In part, the choice of a scaling technique depends on the nature of the continua being scaled. Stevens (1975) pointed out that there are two classes of continua, prothetic and metathetic. A prothetic continuum (e.g. loudness) has degrees of magnitude or quantity and is not amenable to linear partitioning (Stevens, 1974). Studies show that when observers try to partition a prothetic continuum into equal intervals there is a systematic bias to partition the lower end of the continuum into smaller intervals (Berry and

Silverman, 1972; Stevens, 1975). That is, when scaling some perceptual dimensions, listeners do not perceive intervals as equal at different locations on the scale (Stevens, 1971, 1974). The end-result is a set of unequal intervals. A metathetic continuum is a qualitative continuum and has a kind of position (Schiavetti, Metz and Sitler, 1981). Pitch, which varies from high to low, is an example of a qualitative continuum. Stevens reported that listeners are capable of dividing a metathetic continuum into equal intervals. Uncertainty remains as to whether accentedness is prothetic or metathetic and therefore amenable to linear partitioning.

Comparing direction magnitude estimation and a seven-point equal-appearing interval (EAI) scale assists in determining if a continuum is prothetic or metathetic (Stevens, 1975). The procedure involves listeners judging a set of stimuli along the continuum in question (in this case, foreign accent) using interval scaling and direct magnitude estimation. After making judgements about a particular perceptual dimension, the arithmetic mean interval scores are plotted against the geometric mean direct magnitude estimation scores. A linear relationship between the two sets of scaled scores indicates that the continuum is metathetic. If the continuum is metathetic, either interval scaling or magnitude estimation would be appropriate. A curvilinear relationship between interval scores and magnitude estimate scores, on the other hand, would suggest that the continuum scale was prothetic. Only direct magnitude estimation is appropriate if the continuum is prothetic in nature. A systematic research programme is needed to determine which accentedness continua are prothetic and which are metathetic. Such a programme will aid in the appropriate selection of direct magnitude estimation or interval scaling in research and clinical measurement. It is uncertain as to whether all foreign accents are amenable to linear partitioning or not. Identifying the appropriate scale is necessary to proceed in determining the relation between directly measured speech parameters and scaled values of speech.

The aim of this study was to determine empirically whether foreign accentedness (in particular Italian foreign accent in English) is a metathetic or prothetic continuum. Identification of whether perceived foreign accent in Italian speakers of English is prothetic or metathetic in nature would allow application of the most appropriate scaling technique, either interval scaling or direct magnitude estimation. A second aim of this study was to determine the reliability with which listeners could use these scaling techniques. Reliability of the raters was examined because the validity of a scale depends on the listeners' proficiency in performing the scaling task (Gescheider, 1976). Listener proficiency in performing a rating task ratifies the scale used. If listeners cannot use scaling techniques reliably, we cannot obtain good indices of the degree of foreign accent.

Method

Listeners

Two groups of 10 listeners participated in this study. Of the 20 listeners, 14 were female and six were male. Ages ranged from 22 to 47 years (mean = 28.25). All were native monolingual speakers of American English. No subjects reported any history of hearing, speech, or language problems. The listeners who participated in this study knew little about scaling techniques and were not familiar with scaling foreign accent. All listeners reported minimal exposure to Italian-accented English. Each listener was randomly assigned to one of the two groups.

Speech samples

The speech samples were sentences drawn from a larger study by Flege *et al.* (1995). The sentences used were those of six native speakers of English (27–39 years of age) and 90 female Italian speakers. The age of the Italian speakers ranged from 29 to 57 years (mean = 43.7 years). These subjects' ages of arrival in Canada ranged from 1.9 years to 23.3 years (mean = 11.31). Their length of residence in Canada ranged from 14.6 years to 44.3 years (mean = 32).

Two sentences produced by each speaker were used: (1) *The good book was red* and (2) *He turned to the right*. These two sentences were selected at random from the larger corpus of sentences. Two sets of stimuli were developed for each group of listeners. Each group of listeners heard two sentences from three native speakers of English and 45 Italian speakers. The native speakers of English were included to provide insight into the sensitivity of the scale in distinguishing between native and non-native talkers with mild foreign accents. The division of the Italian sentences into two sets was based on previously rated degree of accent. Sentences in the two sets were matched in order to provide listeners with sentences that covered the whole range of accentedness. Previous ratings of the degree of foreign accent of the sentences ranged from 13 to 244 (mean = 154). Each set contained sentences that were considered to reflect a minimal degree of foreign accent and samples that had previously been judged as more strongly foreign accented.

Interval scaling procedures

The interval-scaling task used a seven-point equal-appearing interval scale, where '1' represented the *least accented* sounding sentence and '7' represented the *most accented* sounding sentence. Listeners were encouraged to use any number along the scale if the sample sounded somewhere between least accented and most accented. The listeners were given a short training session before the experimental task, to ensure that they were familiar with the interval scaling procedure.

Direct magnitude estimation

Listeners made numerical estimations of the magnitude of accentedness of each sentence. The first sentence that listeners heard was a standard stimulus (a sentence from an Italian speaker, previously rated as having a foreign accent in the middle of the range of sentences under examination). The word 'modulus' introduced the standard stimulus. Listeners were told to give the standard stimulus a numerical value of 100. Listeners scaled the sentences that followed relative to the modulus. For example, if a listener perceived a sentence to be twice as accented as the modulus a value of 200 would be given. After the tenth sample, the listeners heard the standard stimulus again and gave it the same numerical value given on the first occasion. The modulus was reintroduced after every 10 sentences to prevent difficulty recalling the modulus and causing a shift in the listeners' internal standards of accentedness.

Before beginning the experiment, listeners were familiarized with the DME procedure. The sentences presented in the training session covered the entire range of accentedness. The training speech samples were obtained from Italian speakers not included in this study.

Listening sessions

The two groups of listeners participated in two 30-min sessions. The sessions were approximately 1 week apart. Group 1 scaled accentedness in the first session using DME. In the second session, accentedness was scaled using interval scaling. The order of the scaling techniques was reversed for Group 2.

Listeners were seated alone in a soundproof booth. They heard the recorded speech samples through headphones (Sennheiser, HD 490) at a comfortable listening level. A computer program (WINSPARCS, Smith, 1994) output the speech samples. For the interval-scaling task, a computer monitor displayed seven buttons labelled from 1 to 7. After hearing each speech sample the subjects clicked, with a mouse, the number that corresponded to his/her perceptions of the accentedness of the sentence. Once the subject had recorded a response the next sentence was played out. This procedure was repeated until the subject had responded to all the sentences. All listeners' responses were saved to files for later analysis. The subjects heard each sentence only once in random order. Twenty-four sentences selected randomly were repeated at the end of the sentence set to evaluate reliability.

The same computer program was used for DME. The listeners, after hearing each stimulus, typed in the numerical value that corresponded to his/her perceptions of the magnitude of accentedness of the speech sample. Listeners were told to use any numbers they wished, with the exception of negative numbers. If a listener typed in a negative number it was rejected by the program and he/she was asked to enter another positive value. The computer recorded all responses for later analysis. Stimuli were played out in random order. Twenty-four sentences selected randomly were presented twice to measure reliability.

Results

Intra-judge reliability

Accentedness intra-class correlations (Ebel, 1958) were computed as measures of reliability for DME and interval scores for both groups of listeners (see table 1). Intra-class correlations for interval scaling ranged from 0.78 to 0.97 for Group 1 and from 0.11 to 0.95 for Group 2. All subjects in Group 1 achieved an acceptable level of reliability for interval scaling. A correlation of 0.75 is considered to be acceptable (Shrout and Fleiss, 1979). Seven of the subjects in Group 2 achieved an acceptable level of reliability for interval scaling. DME intra-class correlations ranged from 0.13 to 0.95 for Group 1 and 0.48 and 0.93 for Group 2. DME reliability coefficients were acceptable for eight listeners in Group 1 and nine listeners in Group 2. The poor reliability coefficients for some listeners suggest that a shift may have occurred in their internal standards of accentedness.

Percentage agreement scores were also computed between each listener's original interval scale scores and those of the 24 repeated samples (see table 2). Between 46% and 83% of the samples were given the same score for Group 1. From 0% to 37.5% of the samples differed by ± 1 scale score. Between 25% and 83% of samples

Table 1. *Interval scaling and direct magnitude estimation intra-class correlations for each listener in both groups*

	Group 1		Group 2	
Listeners	Interval scaling	DME	Interval scaling	DME
1	0.87	0.88	0.91	0.83
2	0.95	0.69	0.11	0.86
3	0.89	0.13	0.93	0.92
4	0.95	0.92	0.93	0.92
5	0.97	0.82	0.95	0.93
6	0.90	0.84	0.39	0.48
7	0.90	0.80	0.94	0.76
8	0.78	0.81	0.88	0.86
9	0.87	0.95	0.57	0.80
10	0.88	0.81	0.87	0.82

Table 2. *Listeners' percentage agreement scores for judgements of accentedness that were equal or ± 1 scale score*

	Group 1		Group 2	
Listeners	Equal	± 1 Scale score	Equal	± 1 Scale score
1	50.00	33.33	66.67	33.33
2	83.33	8.33	83.33	8.33
3	79.17	0.00	29.17	70.83
4	62.50	33.33	70.83	12.50
5	66.67	33.33	62.50	37.50
6	54.17	37.50	25.00	41.67
7	45.83	37.50	79.17	12.50
8	70.83	16.67	41.67	41.67
9	58.33	20.83	28.17	29.17
10	54.17	33.33	54.17	37.50

from Group 2 were given the same score. Between 8% and 70% of their scores differed by ± 1 scale score. The agreement scores taken together indicate a high level of reliability. Most of the listeners' scores either equalled the original or differed by only one scale score.

Inter-judge reliability

Intra-class correlations were also computed to assess inter-judge reliability for both groups of listeners. Correlation coefficients were 0.69 and 0.85 for the DME and EAI scores for Group 1 and 0.68 and 0.58 for Group 2. The lower reliability coefficient for each group suggests that listeners' internal standards of accentedness differed.

Frequency distribution of scores

Examining the distribution of DME scores may provide insight into the low reliability coefficients across the listeners. Figure 1 shows the frequency distribution of scores for both groups of listeners. The scores for Group 1 ranged from 2 to 500

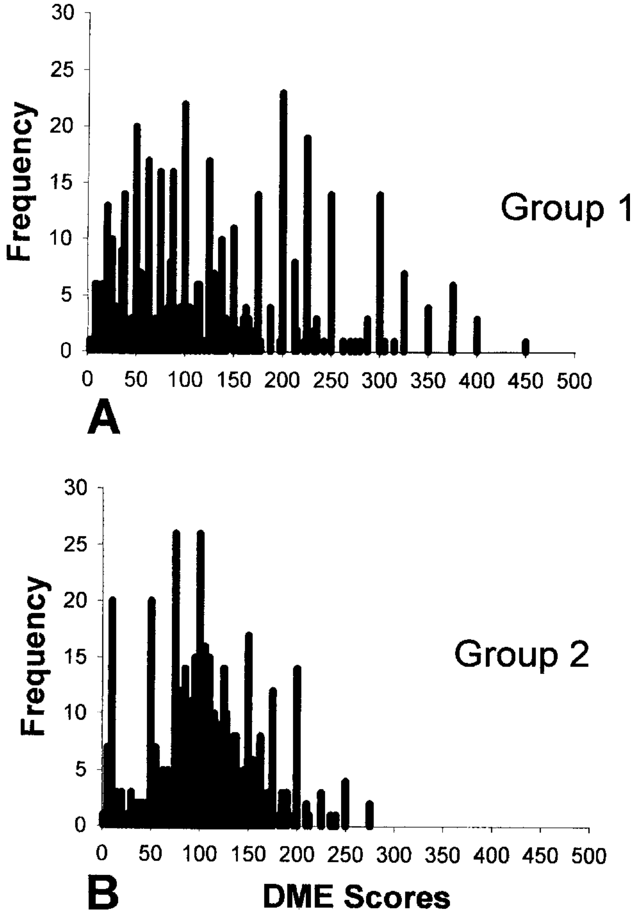

Figure 1. *Frequency distributions of accentedness DME scores for Group 1 (panel A) and Group 2 (panel B).*

(see figure 1a). For Group 2 scores ranged from 1 to 275 (see figure 1b). From the figure we can see that Group 1 used a broader range of DME scores than Group 2. The mode (200) for Group 1 shows they judged the sentences as having stronger foreign accents. Listeners rated sentences 2 to 3 times more accented than the standard sentence. Group 2 rated the sentences more conservatively; their mode was 75. Group 2 more frequently gave ratings of accentedness that were less than the modulus.

The distributions of interval scores were different for both groups (see figure 2). Figure 2a shows that listeners in Group 1 frequently used the anchor points of the scale. Group 2 evenly distributed scores across each interval (see figure 2b). The differences in these distributions suggest that listeners use scales differently. Group 1 provided broad ranges of scores when scaling with DME but tended to use anchor points when using an interval scale. Anchor points may have been used for this group because of possible ceiling effects at the low and high ends of the continuum on the interval scale. DME allowed them to differentiate between accentedness at the low and high ends of the continuum. Group 2 produced an equal distribution

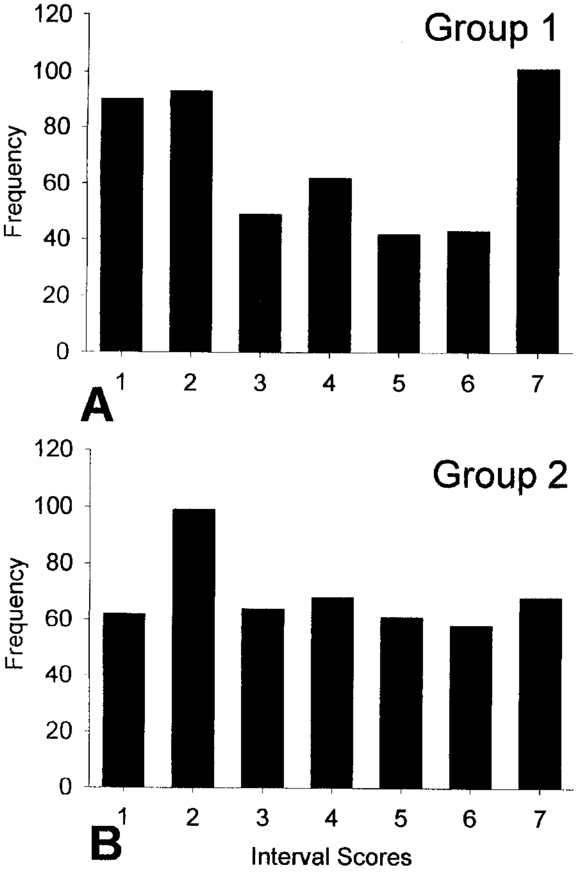

Figure 2. *Frequency distributions of accentedness interval scores for Group 1 (panel A) and Group 2 (panel B).*

of scores across the interval scales and tended to provide DME scores that were closely distributed around the modulus.

Regression analyses

One-way analyses of variance (ANOVA) were computed to reduce the data before computing the regression equations. An ANOVA was computed to determine if differences existed across the scale scores provided by all listeners in each task. For Group 1 and Group 2 significant differences were observed across listeners' interval scale scores ($F_{(9,950)} = 3.15$, $p < 0.01$; $F_{(9,950)} = 5.25$, $p < 0.01$). The DME scores across listeners differed significantly for both groups ($F_{(9,950)} = 10.86$, $p < 0.01$; $F_{(9,950)} = 2.59$, $p < 0.01$). Due to the significant differences in scores across listeners, mean scores were not calculated across listeners' ratings. A further ANOVA was computed to determine if each listener's scores for the two different sentences differed significantly. No significant differences were observed. To reduce the data means were computed across the scale scores for each of the sentences.

The regression results for Group 1 and Group 2 are presented separately.

Group 1. The accentedness DME scores (x-axis) are plotted against the interval scale scores (y-axis) in figure 3a. The individual scores are not tightly clustered around the regression line, particularly at the high end of the continuum. Listeners gave a broad range of DME scores for a particular interval scale score. For example, some samples were consistently given an interval score of 4, but DME scores ranged from 10 to 500. Although scores were not tightly clustered around the regression line the linear relationship between DME and interval scores was significiant ($F_{(1,478)} = 540.45$, $p < 0.001$). Interval scores accounted for 52.96% of the DME variance.

Group 2. Figure 3b shows the relationship between DME and interval scale scores for Group 2. The clustering of the data points around the regression line is tighter than that observed for Group 1; however, some samples given strong accentedness ratings on the interval scale were given mild accentedness ratings with DME. For example, DME scores ranged from 25 to 275 for an interval score of 7. The tighter clustering of scores may be related to the reduced numerical range used by listeners in Group 2. The linear relationship between the interval scores and DME scores

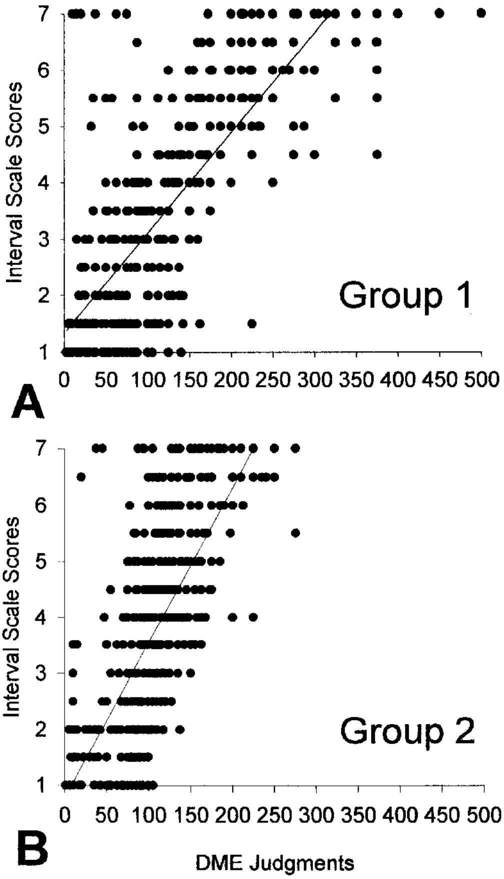

Figure 3. *Scatter plots of interval scores and direct magnitude judgements and best fit regression lines for Group 1 (panel A) and Group 2 (panel B).*

was significant ($F_{(1,478)} = 720.99$, $p < 0.001$). Interval scale scores accounted for 60.04% of the DME variance.

Discussion

The data show a significant linear relationship between interval scores and DME scores. The linear relationship between these two sets of scores suggests that accentedness (at least for Italian speakers of English) is a metathetic continuum. The finding that the relationship between the two sets of scores was linear for two different groups of native English listeners rating two different groups of native Italian speakers strengthens this conclusion. Therefore, an interval scale is appropriate for scaling the accentedness of Italian speakers of English. Although significant, the variance in interval scores accounted for by DME scores was relatively small. Possible response biases caused by the number ranges used by listeners and a potential ceiling effect at the high end of the continuum may have influenced the data and inter-judge reliability.

Response biases come about because perceptual dimensions, such as foreign accent, have no known physical units. Listeners are uncertain how to map responses onto the stimuli because of this lack of known physical units. They have to make their own assumptions about how to map numbers onto stimuli. Because listeners do not have familiar units with which to scale accentedness, they may attempt to use their own units of equal discriminability to partition both continua (see Poulton, 1989). Imposing units of equal discriminability results in a contraction bias in which small differences from the modulus are overestimated and large differences underestimated. The distribution of scores for both groups suggest a contraction bias. Further studies are required to resolve whether the accentedness continuum is truly metathetic or in part the result of contraction biases or listener differences.

The linearity of the function may have in part been due to a scaling artifact such as the logarithmic response bias described by Poulton (1989). This bias results from a tendency for judges to distribute magnitude estimates logarithmically along the number scale, regardless of the true underlying natures of the perceptual continua (metathetic or prothetic). This bias is more likely to occur if the dimension being scaled does not have an easily defined physical referent (Poulton, 1989). A step change in the number of digits used causes a logarithmic response bias (e.g. 1, 20, 25, 35, 50, 100, 200, 300, 400, 500). A large step change occurs at 100. The linearity of the function in this study is partly related to a logarithmic response bias. Some of the listeners produced large step changes at 100. For example, one subject produced changes in ratings in steps of 5 before 100 and changes in ratings in steps of 100 after the modulus.

In order to circumvent the problem of a logarithmic response bias it may be worthwhile to plot such scores against a dimension that has some defined referent (e.g. intelligibility). Other possibilities for circumventing logarithmic response biases are to use number scales in which large step changes do not occur (e.g. number ranges between 100 and 999). Scaling the stimuli in ascending order may also eliminate the influences of a logarithmic response bias.

This logarithmic response bias may have resulted from the DME instructions. In the DME task, listeners judged ratios rather than differences in magnitude. Birnbaum (1980) states that observers perceive and respond to differences in magnitude rather than ratios. Further, Poulton argues that listeners cannot scale ratios of dimensions that do not have familiar stimulus units. Therefore, having listeners rate differences in magnitude rather than ratios may be a way to determine whether the nature of the accentedness continuum is truly metathetic.

The number ranges used by listeners using DME might have influenced the linearity of the function and inter-judge reliability. Some listeners were conservative in their judgements of perceived degree of foreign accent. Their scores were clustered closely around the modulus (i.e. a value of 100). For example, one listener provided scores ranging from 25 to 160 for sentences perceived as having the least and most degree of perceived foreign accent. Other listeners produced a broader distribution of scores. For example, one listener provided scores ranging from 1 to 400. Differences in the number range used by listeners may have influenced inter-judge reliability. It is difficult to determine from this study if the overall shape of the function was due to the different number ranges listeners used or to differences in the listeners' internal standards of what constitutes 'foreign accent'. Therefore, it is necessary to determine if poor inter-judge reliability is a scaling artifact, if it reflects true differences in listeners' internal standards or is a combination of both.

Examination of the raw scores suggests that differences existed among listeners' internal standards. For a particular sentence the listeners provided a broad distribution of DME scores. For example, one particular listener gave a foreign accent rating of 15 for one sentence whereas another listener gave a rating of 250 to the same sentence. The inter-rater differences show that what one listener may perceive as a mild foreign accent another listener may perceive as a strong foreign accent. However, the differences in numbers may reflect listener uncertainty in mapping numbers onto dimensions without defined physical referents (Poulton, 1989).

However, other factors suggest that some of the differences observed across listeners might reflect scaling artifact. One factor that seems to indicate scaling artifact is the number of numbers used by listeners. For example, one listener used only four numbers when scaling degree of perceived foreign accent with DME (50, 100, 200, 300). Another listener used 19 different numbers. The differences in the numbers used suggest that the listeners were uncertain as to how to map numbers onto stimuli.

Further, the frequent use of the anchor points by some listeners when using interval scaling suggests difficulty distinguishing among sentences with similar degrees of foreign accent. Allocation of a particular stimulus to a particular interval on a scale depends on the ease with which that stimulus can be discriminated from another. Easily confused stimuli are placed in the same category. Stimuli most easily distinguishable from each other are placed in different categories. Difficulty discriminating among sentences varying in degree of perceived foreign accent may relate to the lack of definable physical referents that decisively differentiate stimuli.

The frequency distribution data showed that listeners tended to distribute responses equally across the intervals. The tendency for some listeners to distribute scores equally across intervals exemplifies scaling artifact. Using intervals with equal frequency may have influenced the linearity of the function without truly reflecting differences in the degree of perceived foreign accent. A similar problem occurred with DME. The listeners may have attempted to use all responses equally, often causing a centre bias. A centre bias occurs when listeners attempt to use responses above the modulus as often as those below it. A DME centre bias and equal use of intervals would increase the likelihood of a linear function. The contribution of such biases to the shape of the regression function needs further study in order to confirm whether accentedness is a prothetic or metathetic continuum.

The DME procedure requires listeners to judge stimuli relative to a standard stimulus or modulus. The standard stimulus usually comes from the middle of the range of stimuli to avoid skewing the function. However, judging each variable stimulus against a constant standard stimulus may cause a relative stimulus contraction bias (Poulton, 1989). The stimuli are split into two subranges when the standard stimulus lies within the range of stimuli. Thus, the average size of the difference between the standard stimulus and the variable stimulus becomes the reference magnitude for each subrange. Listeners will then underestimate the values of stimuli that differ from the standard by values larger than the average difference.

Both DME and interval scaling require listeners to judge one stimulus after another. In DME the stimulus previous to the one being rated provides an additional reference magnitude against which the current stimulus is judged. Listeners underestimate the size of the difference between the previous stimulus and the next stimulus. The previous stimulus becomes the reference magnitude. The response to the next stimulus becomes the difference in magnitude between the two stimuli. Stimuli larger than the previous stimulus are underestimated and variables smaller than the previous stimulus are overestimated. Pairing the standard stimulus with each variable stimulus may overcome sequential contraction biases.

Examination of the two groups' scale scores suggests a ceiling effect at the high end of the continuum. Listeners gave sentences that were differentiated by DME scores the same score of 7 on the interval scale. The ceiling effect that occurred for foreign accent may be due to the number of scale intervals. A seven-point scale, although frequently used, may not be sensitive enough for all listeners to discriminate among sentences examined here. An 11- or nine-point scale might improve listener sensitivity when scaling degree of perceived foreign accent.

To indirectly assess whether a larger number of intervals might improve scaling resolution, we conducted a *post-hoc* analysis of accentedness ratings obtained by Flege *et al.* (1995) for the subset of native Italian speakers examined in the present study. In the earlier study the native Italian subjects were assigned to one of 10 subgroups based on their age of arrival (AOA) to Canada from Italy. The perceived degree of foreign accent increased linearly as AOA increased. Accentedness ratings of the native Italians speakers' productions of the two sentences considered here, differed significantly from the native English (NE) controls.

We asked the question whether the same pattern of significant between-group differences obtained earlier using continuous scale ratings would be obtained if the scores were transformed into five-, seven-, nine-, 11- and 13-point interval scales. The continuous scale contained values ranging from 0 to 255. To convert the continuous scale into an interval scale it was arithmetically subdivided into number ranges equivalent to the number of intervals on each scale. For example, to convert the continuous scale into a five-point scale, the continuous scale was subdivided into the following five equal ranges: 0–51, 52–102, 103–153, 154–204, 205–255. The original data were then assigned a value of 1 to 5 based on the subrange into which it fell (e.g. a value of 31 was coded as '5', a value of 100 was coded as '2', and so on). The same mathematical procedure was used to convert the continuous scale into the other interval scales.

Transformed ratings on the nine-, 11-, and 13-point scales yielded significant differences between accentedness ratings of NE controls and native Italian speakers. Analysis of the five- and seven-point scale data for the sentence, '*He turned to the right*' failed to yield a significant difference between the NE control accentedness ratings and those for the native Italian speakers. For '*The good book was red*', only the 5-point scale data failed to yield a significant between-group difference. This data suggest that a seven-point scale may not allow NE listeners to rate adequately variations in perceived degree of Italian-accented English.

The lack of a physical referent may increase variability because listeners do not use the same acoustic cues to form the basis of their accentedness judgements. In addition to the variance in the use of acoustic cues across listeners, the number of acoustic cues used by listeners to judge degree of perceived foreign accent may vary. Some listeners may rely on a number of different acoustic cues whereas others may rely on only one or two acoustic cues. This may in turn affect listeners' perceptions of overall degree of foreign accent. Listeners who attend to many acoustic cues may give stronger foreign accent ratings than listeners who rely on only one or two acoustic cues. Identification of the potential acoustic cues used by listeners may help provide a physical referent to assist in interpreting judgements of degree of perceived foreign accent.

Uncertainty about the perceptual dimensions that underlie foreign accent may have also created the broader distribution of scores at the high end of the accentedness continuum. A large number of perceptual dimensions may create the perception of a strong foreign accent. Therefore, listeners may be uncertain as to which potential perceptual dimensions influence their judgements. When judging strong accents, listeners have to concentrate on a number of perceptual dimensions. Therefore, the more distant the non-native sentence is from the native speaker norm, the more likely it is that the criteria listeners use will differ. Faust (1984) argues that listeners attend to only a few cues when making judgements about behavioural dimensions. Identifying the perceptual dimensions underlying accentedness may assist in providing familiar units that listeners can use to judge foreign accent.

Conclusions

Overall, the findings from this study suggest that, at least when judging Italian native speakers, English-speaking listeners are capable of partitioning accentedness into equal intervals. The data suggest that accentedness may be a metathetic continuum. Either an equal-appearing interval scale or DME can provide valid indices of accentedness. Due to the possible influences of scaling artifact, the nature of the accentedness continuum requires further examination. Response bias effects appear related to the lack of definable physical referents for the accentedness dimension, at least in sentences spoken with a mild-to-moderate foreign accent. Although response biases may have contributed to the linearity of the function, they cannot entirely explain the results. The replication of results across two groups of listeners scaling different sentences confirms the likelihood that for Italian accentedness the continuum is metathetic.

The response biases outlined obviously affected inter-judge reliability. The lack of familiar units forced listeners to make their own assumptions about how to map responses onto the sentences, resulting in differences in the listeners' internal standards. Exposure and familiarity with foreign accent may cause differing internal standards, ultimately affecting the validation of acoustic cues through scaling. Experienced listeners who are familiar with foreign accent may provide lower accentedness ratings than listeners who have had little exposure. Further studies are required to determine how familiarity and experience with foreign accent influence the nature of the accentedness continuum.

The finding that accentedness is a metathetic continuum only pertains to native Italian speakers of English. From this study, generalizations about the nature of the accentedness continua of other non-native speakers of English or speakers with stronger foreign accents are inappropriate. Further research is necessary to determine whether DME or EAI is appropriate to scale other foreign (or stronger) accents.

This study highlights that scaling perceptual dimensions without definable physical units is fraught with many problems. Further research is required to investigate scaling techniques in order to determine if observed differences among listeners reflect differences in their internal standards or relate in part to scaling artifact. Identification of requisite perceptual dimensions for scaling accentedness may explain differences in listeners' internal standards. Pursuing research that relates acoustic variables to global judgements of degree of perceived foreign accent would be profitable.

Acknowledgements

This study was supported in part by grant RO1-DC00257 from the National Institute for Deafness and Other Communicative Disorders. We thank Gwyneth Cliver for assistance with data collection and analysis. Thanks also go to Stephen C. Smith for design and programming of the software package used in this experiment. We thank two anonymous reviewers for their helpful comments on a previous draft of this manuscript.

References

Asher, J. and Garcia, R., 1969, The optimal age to learn a foreign language. *Modern Language Journal*, **53**, 334–341.

Berry, R. C. and Silverman, F. H., 1972, Equality of intervals on the Lewis-Sherman Scale of Stuttering Severity. *Journal of Speech and Hearing Research*, **15**, 185–188.

Birnbaum, M. H., 1980, Comparison of two theories of 'ratio' and 'difference' judgments. *Journal of Experimental Psychology: General*, **109**, 304–319.

Brennan, E. and Brennan, J., 1981, Accent scaling and language attitudes: reactions to Mexican-American English speech. *Language and Speech*, **24**, 207–221.

Brennan, E., Ryan, E. and Dawson, W., 1975, Scaling of apparent accentedness by magnitude estimation and sensory modality matching. *Journal of Psycholinguistic Research*, **4**, 27–36.

Ebel, R. L., 1958, Estimation of the reliability of ratings. *Psychometrika*, **16**, 407–424.

Faust, D., 1984, *The Limits of Scientific Reasoning* (Minneapolis, MN: University of Minnesota Press).

Flege, J. E., 1984, The detection of French accent by American listeners. *Journal of the Acoustical Society of America*, **76**, 692–707.

Flege, J. E., 1988, Factors affecting degree of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, **84**, 70–79.

Flege, J. E. and Fletcher, K., 1992, Talker and listener effects on the perception of the degree of foreign accent. *Journal of the Acoustical Society of America*, **91**, 370–389.

Flege, J. E. and Hillenbrand, J., 1984, Limits on pronunciation accuracy in adult foreign language speech production. *Journal of the Acoustical Society of America*, **76**, 708–721.

Flege, J. E., Munro, M. J. and MacKay, I., 1995, Factors effecting the degree of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, **97**, 3125–3134.

Gescheider, G. A., 1976, *Psychophysics, Method and Theory* (Hillsdale, NJ: Lawrence Erlbaum).

Mack, M., 1982, Voice-dependent vowel duration in English and French: monolingual and bilingual production. *Journal of the Acoustical Society of America*, **71**, 173–178.

Munro, M. J., 1993, Productions of English vowels by native speakers of Arabic: acoustic measurements and accentedness ratings. *Language and Speech*, **36**, 39–66.

Munro, M. J., 1998, The effects of noise on the intelligibility of foreign-accented speech. *Studies in Second Language Acquisition*, **20**, 139–154

Oyama, S., 1982, The sensitive period for the acquisition of a non-native phonological system. In S. Krashen, R. Scarcella, and M. Long (Eds), *Child–Adult Differences in Second Language Acquisiton* (Rowley, MA: Newbury House), pp. 20–38.

Poulton, E. C., 1989, *Bias in Quantifying Judgment* (Hillsdale, NJ: Lawrence Erlbaum).

Purcell, E. and Suter, R., 1980, Predictors of pronunciation accuracy: a reexamination. *Language Learning*, **30**, 271–287.

Ryan, E., Carranza, M. and Moffie, R., 1977, Reactions towards varying degrees of accentedness in the speech of Spanish-English bilinguals. *Language and Speech*, **20**, 267–273.

Schiavetti, N., Metz, D. E. and Sitler, R. W., 1981, Construct validity of direct magnitude estimation and interval scaling: evidence from a study of the hearing-impaired. *Journal of Speech and Hearing Research*, **24**, 441–445.

Sebastian, R., Ryan, E., Keogh, T. and Schmidt, A., 1980, The effects of negative affect arousal on reactions to speakers. In H. Giles, W. Robinson, and P. Smith (Eds), *Language: Social Psychological Perspectives* (Oxford: Pergamon Press), pp. 195–213.

Shrout, P. E. and Fleiss, J. L., 1979, Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, **86**, 420–428.

Smith, S. C., 1994, SPARCS for Windows. [Computer Software]. Birmingham, AL. Department of Biocommunication, University of Alabama at Birmingham.

Stevens, S. S., 1971, Issues in psychological measurement. *Psychological Review*, **78**, 426–450.

Stevens, S. S., 1974, Perceptual magnitude and its measurement. In E. C. Caterette and M. P. Friedman (Eds), *Handbook of Perception*, Vol. 2 (New York: Academic Press).

Stevens, S. S., 1975, *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects* (New York: John Wiley).

Williams, L., 1980, Phonetic variation as a function of second-language learning. In G. Yeni-Komishian, J. Kavanagh and C. Ferguson (Eds), *Child Phonology*, Vol. 2: *Perception* (New York; Academic Press), pp. 185–216.

Young, M. A., 1969, Observer agreement: cumulative effects of rating many scales. *Journal of Speech and Hearing Research*, **12**, 135–143.