SSLA, **27**, 597–606. Printed in the United States of America.
DOI: 10.1017/S0272263105050266

AN APPROPRIATE METRIC FOR CUE WEIGHTING IN L2 SPEECH PERCEPTION

Response to Escudero and Boersma (2004)

Geoffrey Stewart Morrison
University of Alberta

Flege, Bohn, and Jang (1997) and Escudero and Boersma (2004) analyzed first language-Spanish second language-English listeners' perception of English /i/–/ɪ/ continua that varied in spectral and duration properties. They compared individuals and groups on the basis of spectral reliance and duration reliance measures. These reliance measures indicate the change in identification rates from one extreme of the stimulus space to the other; they make use of only a portion of the data collected and suffer from a ceiling effect. The current paper presents a reanalysis of Escudero and Boersma's data using first-order logistic regression modeling. All of the available data contribute to the calculation of logistic regression coefficients, and they do not suffer from the same ceiling effect as the reliance measures. It is argued that—as a metric of cue weighting—logistic regression coefficients offer methodological and substantive advantages over the reliance measures.

In a recent study, Escudero and Boersma (2004) analyzed first language (L1)-Spanish second language (L2)-English listeners' perception of a two-dimensional

My thanks to Paola Escudero and Paul Boersma for making their data available, and thanks to Terrance M. Nearey for comments on an earlier draft of this paper (any defects are my own responsibility). This work was supported by the Social Sciences and Humanities Research Council of Canada.

Address correspondence to: Geoffrey Stewart Morrison, Department of Linguistics, 4-32 Assiniboia Hall, University of Alberta, Edmonton, Alberta, T6G 2E7, Canada; e-mail: gsm2@ualberta.ca

© 2005 Cambridge University Press 0272-2631/05 $12.00

English /i/–/ɪ/ continuum that varied in spectral and duration properties. They quantified the listeners' response data using spectral reliance and duration reliance measures, a metric that had previously been used by Bohn (1995) and Flege, Bohn, and Jang (1997) (and referred to in the latter studies as *spectral effect* and *temporal effect* scores). The use of the same metric by different groups of researchers suggests that it might be on its way to becoming a standard measure of cue weighting in L2 speech perception research. I propose, instead, that logistic regression coefficients provide a more appropriate metric and should be adopted in place of reliance measures. In this paper, I will first describe the reliance metric and some of its disadvantages. I will then describe the logistic regression coefficient metric and some of its advantages. Examples of predicted probability plots based on logistic regression analyses of several listeners' data will be discussed, along with a comparison of the logistic regression coefficients and reliance measures for the same listeners. Following this, I will compare the distributions of the reliance measures and logistic regression coefficients and show that the latter are more appropriate for use as dependent variables in statistical tests. Finally, I will conduct statistical tests on logistic regression coefficients in order to answer research questions from Escudero and Boersma.

To calculate the duration reliance, one calculates the proportion of /i/ responses for the set of stimuli with the longest duration (pooled across F1) and then subtracts the proportion of /i/ responses for the set of stimuli with the shortest duration. In a two-dimensional grid with duration on the horizontal axis and F1 on the vertical axis (see Figure 1 from Escudero & Boersma, 2004), the proportion of /i/ responses for the extreme left-hand column is subtracted from the proportion of /i/ responses for the extreme right-hand column. Similarly, the spectral reliance is the proportion of /i/ responses for the set of stimuli with the lowest F1, less the proportion of /i/ responses for the set of stimuli with the highest F1. This is a crude measure of the overall change in response from one extreme of each dimension to the other, and its calculation made use of only a fraction of the data collected in Flege et al. (1997) and in Escudero and Boersma: As Escudero and Boersma themselves pointed out, calculating each of the spectral and duration reliances made use of data from only 14 of their 37 stimuli. If the unused data were irrelevant, then it would be more efficient not to collect these data in the first place; however, if the unused data are relevant, then it would be preferable to use a metric that took into consideration all of the data collected. The reliance metric reaches a ceiling once there is a 100% change in vowel identification from one end of the dimension to the other and gives no indication of the nature of the response pattern between the dimension extremes: The proportion of /i/ responses could increase linearly across the continuum; it could stay at zero until part way across the continuum, then suddenly change to one; or it could have some sigmoidal shape intermediate between these two. Note that the ceiling is due to the manner in which the reliance metric is calculated and is not due to the nature of the raw response data or experimental design.

An alternative metric that does not suffer from the inadequacies of the reliance metric is provided by logistic regression (see Fleiss, Levin, & Paik, 2003; Hosmer & Lemeshow, 2000; Menard, 2002; Pampel, 2000). Other metrics with similar properties—such as probit coefficients—would suffice, but logistic regression is now a standard and widely used method of analyzing proportional data. Logistic regression has been successfully applied to L1 speech perception data from experimental designs more complex than those of Flege et al. (1997) and Escudero and Boersma (2004) (e.g., Benkí, 2001; Nearey, 1990, 1997), and analyses of dichotomous (binomial) data can easily be conducted using commercially available software such as SPSS or STATA or free software such as R. Forced-choice identification experiments result in proportional data. For example, if a listener hears a particular stimulus 10 times and responds /i/ on 8 occasions and /ɪ/ on 2 occasions, then the proportion of /i/ responses is 8/10 = 0.8. Whatever the combination of /i/ and /ɪ/ responses, the total must add up to 10 because the listener heard the stimuli 10 times and was obliged to provide a response on each occasion. Proportional response data have values that vary from 0 to 1; because of this range limit, proportional data are inappropriate for linear regression analysis. In the extreme case, analyzing proportional data using linear regression can lead to nonsensical answers such as the prediction that a listener presented with 10 stimuli would give 12 /i/ responses. To avoid range-limit problems, logistic regression operates in a logistic (or logged odds) space rather than a probability space. The model's predicted values in logits can subsequently be converted to probabilities.[1] Logistic regression fits a model to response data using an iterative maximum likelihood technique to derive estimates of coefficients in equations of the form $Logit(/\text{i}/ \mid dur, spec) = \alpha + \beta_{\text{dur}} \times dur + \beta_{\text{spec}} \times spec$, where *dur* and *spec* are stimulus property values for duration and spectral properties, respectively, α is a bias coefficient that is independent of changes in stimulus properties, and β_{dur} and β_{spec} are coefficients that represent the tuning of the response by the stimulus properties. The stimulus-tuned coefficients indicate how fast the response changes from /ɪ/ to /i/ as the duration increases and F1 decreases. They indicate slopes of straight lines (a two-dimensional flat plane) in the logistic space, which are related to the maximum slope of a tangent to the sigmoidal lines (the two-dimensional sigmoidal surface) in the proportion space.[2] The coefficient values have a practical ceiling at the slope of a 100% change in response between two adjacent rows or columns of the continuum, as opposed to a 100% change from the two most distant rows or columns for the reliance measures. Reaching a ceiling in logistic regression coefficients usually indicates that the continuum sampled the acoustic space too coarsely.

Figures 1–4 provide sample probability surface plots based on logistic regression analyses of individual participants' data. The height of the darker surface above the base of the plot represents the predicted probability of an /i/ response, and the height of the lighter surface represents the predicted probability of an /ɪ/ response. These are smoothed representations of the lis-

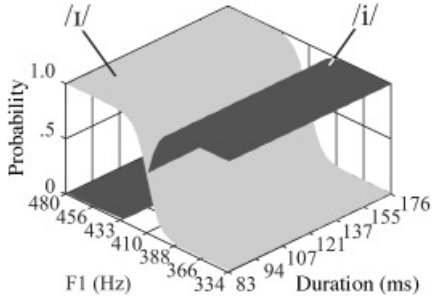

Figure 1. Sample three-dimensional probability plot based on logistic regression analyses of L1-Scottish English participant LH's response data from Escudero and Boersma (2004).

teners' response data. Figure 1 is the plot based on an analysis of the response data from one of the L1-Scottish English listeners (LH), which resulted in coefficient values of $\alpha = -8.75$, $\beta_{dur} = 0.01$, and $\beta_{spec} = 2.21$, and the plot in Figure 2 is based on an analysis of the response data from one of the L1-Southern English listeners (RH), which resulted in coefficient values of $\alpha = -4.35$, $\beta_{dur} = 0.30$, and $\beta_{spec} = 0.83$.[3] In both plots, as F1 decreases, the predicted probability of /i/ responses increases; however, the rate of change from /ɪ/ to /i/ responses—the slopes of the planes in the spectral dimension—is much steeper for LH than for RH. Participant LH therefore has a greater spectral cue weighting than participant RH, as reflected in the difference in the values of the spectrally tuned logistic regression coefficient β_{spec}: 2.21 versus 0.83. In comparison, the spectral reliance measurements for participants LH and RH (1.00 vs. 0.91) suggest much less of a difference in spectral cue weighting; this is because the spectral reliance for LH is at the ceiling value even though

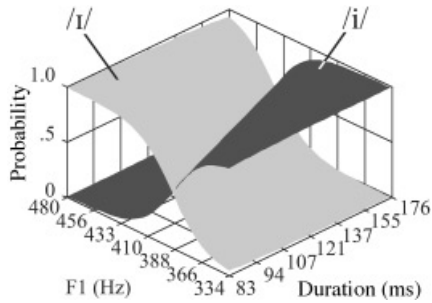

Figure 2. Sample three-dimensional probability plot based on logistic regression analyses of L1-Southern English participant RH's response data from Escudero and Boersma (2004).

the pattern of change from /ɪ/ to /i/ responses far exceeds the minimum necessary to reach the ceiling value. The plot for LH (Figure 1) has the probability of an /ɪ/ response at or close to 100% for the three highest F1 values (480–433 Hz) and shows the probability of an /i/ response at or close to 100% for the three lowest F1 values (388–344 Hz), with a sudden jump around 410 Hz. The minimum pattern necessary to reach ceiling—a gradual shift from 100% /ɪ/ responses for stimuli with the highest F1 (480 Hz) to 100% /i/ responses for stimuli with the lowest F1 (344 Hz)—would look more similar to the plot for RH (Figure 2), whose spectral reliance was just below ceiling.

Turning to the duration dimension, the plots suggest that LH did not make use of duration cues, whereas RH made some use of these cues; for RH, longer vowels have a higher predicted probability of /i/ responses. The difference in duration cue weighting is reflected in the values of the duration-tuned logistic regression coefficient β_{dur} for LH and RH (0.01 vs. 0.30), which were similar to the duration reliance measures, −0.03 versus 0.30. Note that although LH's duration-tuned logistic regression coefficient suggests a very slight positive global trend in the relationship between vowel duration and /i/ responses, the duration reliance measure indicated a very slight trend in the opposite direction when only stimuli with the most extreme durations are considered. An additional advantage of logistic regression coefficients is that various tests of significance are available: A Wald test on LH's β_{dur} value indicated that it was not significantly different from zero, $\chi^2 = 0.021$, $p = .884$. In other words, vowel duration did not affect this listener's choice of /i/ or /ɪ/ responses. (Unless otherwise indicated, all other coefficient values quoted were significant at an α level of 0.001.) The orientation of the boundary line between predominantly /ɪ/ responses and predominantly /i/ responses is specified by the ratio of β_{dur} to β_{spec}.[4] This boundary line is the intersection of the two surfaces, which is at .5 predicted probability for each phoneme.

Figures 3 and 4 are based on analyses of response data from two L1-Spanish L2-Southern English listeners (LJ and MC), which resulted in coefficient val-

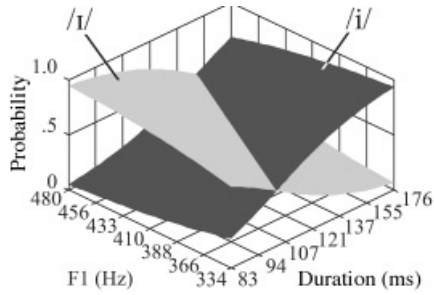

Figure 3. Sample three-dimensional probability plot based on logistic regression analyses of L1-Spanish participant LJ's response data from Escudero and Boersma (2004).

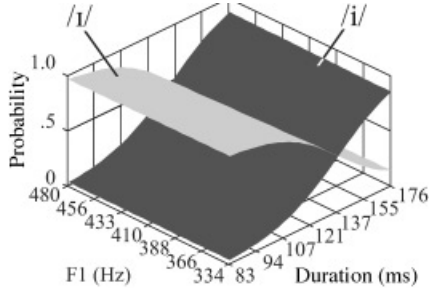

Figure 4. Sample three-dimensional probability plot based on logistic regression analyses of L1-Spanish participant MC's response data from Escudero and Boersma (2004).

ues of $\alpha = -1.92$, $\beta_{dur} = 0.31$, and $\beta_{spec} = 0.16$, and $\alpha = -2.22$, $\beta_{dur} = 0.46$, and $\beta_{spec} = -0.02$, respectively (the value of the latter β_{spec} was not statistically significant, Wald $\chi^2 = 0.217$, $p = .641$). The duration and spectral reliance measures for LJ were 0.70 and 0.36, respectively, and the same measures were 0.79 and −0.03 for MC, respectively. In contrast to the L1-English listeners, these two L1-Spanish listeners made relatively greater use of duration cues and less use of spectral cues.

Figures 5 and 6 provide boxplots of spectral and duration reliance measures and logistic regression coefficients based on analyses of individual participants' perception data. The boxplots indicate that in comparison to the reliance measures, the logistic regression coefficients are closer to having Normal distributions, they are less skewed, and they also have less between-group heterogeneity of variance. The reliance measures suffer from a ceiling effect: Nine L1-Scottish English listeners, one L1-Southern English listener, and two L1-Spanish L2-Scottish English listeners had spectral reliances at the maximum value of 1. The reliance measures, therefore, violate not only the parametric test assumptions of normality but also the assumption for non-parametric ranked data tests (such as the Mann-Whitney *U*-test), which states that the measurement scale be at least ordinal: At the ceiling, the scale ceases to be ordinal, and if there are multiple cases from different groups at the ceiling value, then these cases cannot be ranked relative to one another. Logistic regression coefficient values are therefore more appropriate for use as dependent variables in subsequent statistical tests.

Escudero and Boersma (2004) divided their participants into six groups (listeners who identified /i/ and /ɪ/ using exclusively duration, mainly duration, duration and spectrum, spectrum and duration, mainly spectrum, and exclusively spectrum) on the basis of the ratio of their duration to spectral reliance scores and then conducted one-tailed two-sample Kolmogorov-Smirnov tests on the number of L1-Scottish English, L1-Southern English, and L1-Spanish participants in each group. Although, at first glance, the basis for dividing

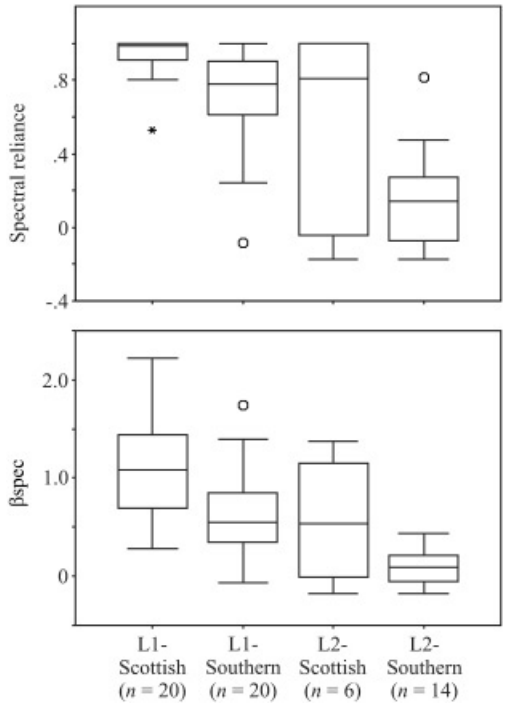

Figure 5. Boxplot of spectral reliance measures and logistic regression coefficients based on analyses of individual listeners' response data from Escudero and Boersma (2004).

participants into groups (duration to spectral reliance ratios of 4 or above, 2–4, 1–2, 1/2–1, 1/4–1/2, and 1/4 or less) appears logical, it is in fact arbitrary unless one assumes that a given change in the rate of identification from one extreme of the duration dimension to the other (83–176 ms) is equally as meaningful as the same change in the rate of identification from one extreme of the spectral dimension to the other (F1: 480–344 Hz and F2: 1893–2320 Hz). Comparing the groups in this manner masks—but does not truly solve—the ceiling effect problem: Some of the ratios are based on spectral reliances that were at ceiling, and it is not conceptually valid to compare these with ratios based on spectral and duration reliances that were not at ceiling. This way of analyzing the data also deviates from standard practice: If one has a set of data with one value per participant, the normal and straightforward procedure is to run a test based on that set of data—not to divide the data into subgroups and then conduct a test. Dividing the participants into groups reduced the sample size (e.g., from 20 to 6 for each L1-English group), which is likely to result in a less powerful test.

L1-Scottish English speakers produce a relatively large spectral difference but no or little duration difference between /i/ and /ɪ/. L1-Southern English

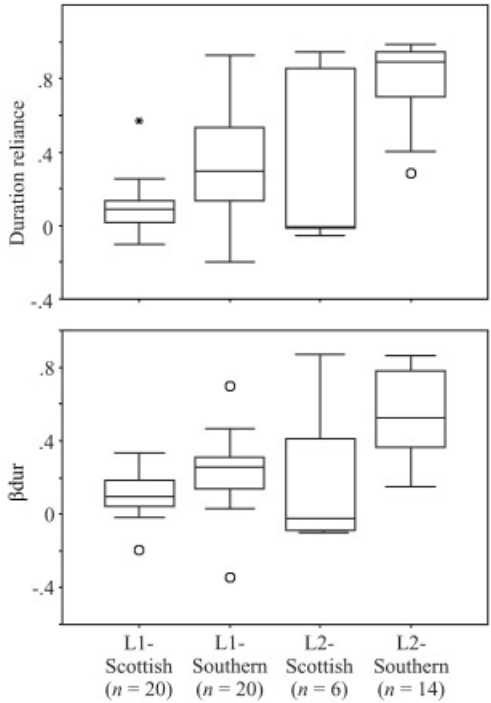

Figure 6. Boxplot of duration reliance measures and logistic regression coefficients based on analyses of individual listeners' response data from Escudero and Boersma (2004).

speakers produce a smaller spectral difference but a large duration difference. In terms of Best's (1995) perceptual assimilation model, Escudero and Boersma (2004) predicted that Scottish /i/ and /ɪ/ would undergo a two-category assimilation to Spanish /i/ and /e/ and that L1-Spanish L2-Scottish English learners would therefore distinguish the /i/–/ɪ/ continuum primarily on the basis of spectral cues. Southern English /i/ and /ɪ/ were predicted to undergo a category-goodness assimilation to Spanish /i/, and L1-Spanish L2-Southern English learners were expected to distinguish the /i/–/ɪ/ continuum primarily on the basis of duration cues. To test these hypotheses, I conducted Welch's *t*-tests (equality of variances not assumed) on the stimulus-tuned logistic regression coefficients. Reflecting differences reported for production, L1-Scottish English listeners made significantly more use of spectral cues than L1-Southern English listeners, $t(36.235) = 3.287$, $p < .01$, and less use of duration cues, $t(30.043) = -2.412$, $p < .05$. L1-Spanish L2-Scottish English listeners made significantly less use of spectral cues than L1-Scottish English listeners, $t(7.431) = -2.005$, $p < .1$, but there was no significant difference in their use of duration cues, $t(5.274) = -0.422$, $p > .690$.[5] L1-Spanish L2-Southern

English listeners made significantly less use of spectral cues than L1-Southern English listeners, $t(27.858) = -4.742$, $p < .001$, and significantly more use of duration cues, $t(26.589) = 3.951$, $p < .01$. L1-Spanish L2-Scottish English listeners did not make substantially more use of spectral cues than L1-Spanish L2-Southern English listeners, $t(5.407) = 1.791$, $p > .129$, but did make substantially less use of duration cues, $t(6.469) = -2.117$, $p < .1$. With respect to L1-Spanish L2-Southern English listeners, the results of the Welch's *t*-tests on logistic regression coefficients were consistent with the predictions: The L1-Spanish L2-Southern English listeners made more use of duration cues than both L1-Southern English and L1-Spanish L2-Scottish English listeners and less use of spectral cues than L1-Southern English listeners. This is consistent with the L2-Southern English listeners—but not L2-Scottish English listeners—making a category-goodness assimilation of English /ɪ/ and /i/ to Spanish /i/ and distinguishing the two English vowels on the basis of their duration differences. With respect to L1-Spanish L2-Scottish English listeners, the results were not consistent with predictions: L1-Spanish L2-Scottish English listeners made less use of spectral cues than L1-Scottish English listeners, and although there was a trend in the expected direction (see the boxplots in Figures 5 and 6), L2-Scottish English listeners did not make significantly more use of spectral cues than L2-Southern English listeners. Both of these results are inconsistent with the prediction that L2-Scottish English listeners would make a two-category assimilation of English /ɪ/ and /i/ to Spanish /i/ and /e/ and distinguish the two English vowels on the basis of their spectral differences. However, perception data from substantially more than six L1-Spanish L2-Scottish English listeners would be required to fully test this hypothesis.

I have argued that logistic regression coefficients are a better metric for cue weighting than the reliance measures of Flege et al. (1997) and Escudero and Boersma (2004): Whereas logistic regression makes use of all data collected, reliance measures are based only on data from the extremes of the stimulus space. Reliance measures also have a ceiling; they cannot distinguish among participants who gradually shift from 100% /ɪ/ responses at one extreme of the stimulus space to 100% /i/ responses at the other extreme, participants who have 100% /ɪ/ responses over half the stimulus space and 100% /i/ responses over the other half and a sudden jump in between, and participants with any pattern between these two. Because a proportion of the cases in Escudero and Boersma's spectral reliance data had values at the ceiling, the distribution of the data did not meet the assumptions for either parametric tests or nonparametric ranked data tests. Logistic regression coefficients do not suffer from the same ceiling as the reliance measures, and boxplots indicated that the data generally met the assumptions for parametric Welch's *t*-tests (although the distribution of the L1-Spanish L2-Scottish English listeners' duration-tuned coefficients was problematic, this can be attributed to the small sample size rather than to the logistic regression coefficient metric). Logistic regression coefficients—but not reliance measures—could therefore be used as the dependent variable in straightforward tests of the between-

group contrasts of interest. Given these factors, logistic regression coefficients provide a superior metric of stimulus weighting and should be adopted in L2 perception research in preference to the reliance measures used in Flege et al. and Escudero and Boersma.

(*Received 17 February 2005*)

NOTES

1. The formula for conversion from a logit to a probability for given duration and spectral values is $prob(/i/ \mid dur, spec) = exp(Logit(/i/ \mid dur, spec))/(exp(Logit(/i/ \mid dur, spec)) + exp(Logit(/ɪ/ \mid dur, spec)))$, which in the binomial case can be simplified to $prob(/i/ \mid dur, spec) = 1/(1 + exp(-Logit(/i/ \mid dur, spec)))$.

2. In the binomial case, using deviation-from-mean coding, the estimated rate of change from /ɪ/ to /i/ in logits is twice the estimated stimulus-tuned logistic regression coefficient value. For example, if the estimated β_{spec} coefficient is 2.207, the slope in the logistic space is 4.414 logits per one step increase in spectral units. In Escudero and Boersma's (2004) stimulus space, a one-step increase in the spectral dimension is a decrease of 23 mel in F1 and covarying increase of 34 mel in F2; a one-step increase in the duration dimension is a 13% increase in duration. The maximum slope of a tangent to the surface in the probability space is one-quarter of the value of the slope in the logistic regression space (see Pampel, 2000, p. 25). For example, if the estimated β_{spec} coefficient is 2.207, the slope of the tangent to the steepest part of the curve in the probability space is an increase in probability of /i/ response relative to /ɪ/ of 1.104 per one-step increase in spectral units, which equals 0.048 per 1-mel decrease in F1.

3. Unless one has some external motivation for believing that a one-unit increase in one dimension is equally as meaningful as a one-unit increase in the other dimension, one should be cautious about making comparisons between spectrally tuned and duration-tuned coefficients or between spectral and duration reliances.

4. The formula for this line is $\alpha + \beta_{dur} \times dur + \beta_{spec} \times spec = Logit(0.5) = log(0.5/0.5) = log(1) = 0$.

5. Given the small number of participants in the L1-Spanish L2-Scottish English group, tests including this group would not be expected to have very much power, and the α level was relaxed to 0.1. The extreme skewedness of the β_{dur} distribution—see Figure 6—makes it unlikely that the reported *p*-values are accurate.

REFERENCES

Benkí, J. R. (2001). Place of articulation and first formant transition pattern both affect perception of voicing in English. *Journal of Phonetics*, *29*, 1–22.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience*: *Issues in cross-language research* (pp. 171–204). Timonium, MD: York Press.

Bohn, O.-S. (1995). Cross-language speech perception in adults: First language transfer doesn't tell it all. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 279–304). Timonium, MD: York Press.

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*, 551–585.

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on nonnative speakers' production and perception of English vowels. *Journal of Phonetics*, *25*, 437–470.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed). New York: Wiley.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Menard, S. (2002). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.

Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, *18*, 347–373.

Nearey, T. M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, *101*, 3241–3254.

Pampel, F. C. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage.