# Measuring relative cue weighting: a reply to Morrison

**by Paul Boersma and Paola Escudero, University of Amsterdam**

**April 28, 2005**

Morrison (2005) criticizes the analytical and statistical methods that Escudero & Boersma (2004), henceforth E&B, used for assessing the cue weightings of the participants in their listening experiments. He proposes that logistic regression constitutes a better method for measuring perceptual cue weighting than E&B's 'edge difference ratio'. The present paper starts by summarizing and illustrating E&B's experiment and analysis method, then addresses five of Morrison's objections, namely the alleged 'ceiling effect', the alleged superiority of logistic regression, the problem of discarding data, the (dis)confirmation of two-category assimilation, and E&B's grouping of the data. We will argue that although logistic regression is a very good method for measuring cue weighting, there was nothing wrong with E&B's methodology in these five respects.

## 1. Escudero & Boersma's listening experiment

In their listening experiments, E&B presented the participants with the 37 different stimuli in Figure 1. The stimuli were synthesized vowel-like sounds that varied in duration (7 possible values) as well as in spectral quality (7 possible values).

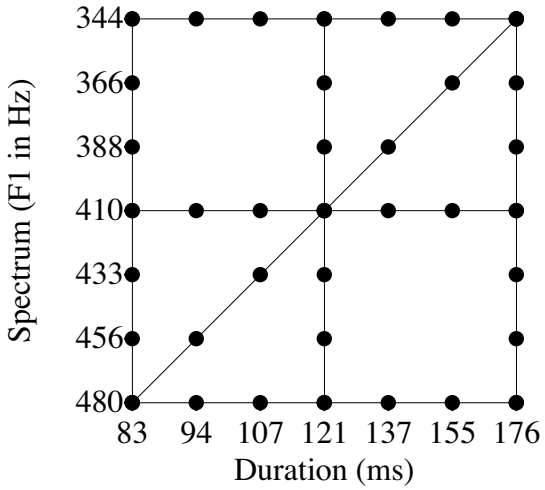


**Fig. 1.** Stimulus set used by Escudero & Boersma (2004).

In the experiment, each of the stimuli was presented 10 times. The 370 trials were presented in a randomized order. The task of the participants was to decide whether the vowel they heard was the English /i/ (by clicking on a picture of a *sheep*) or the English /ɪ/ (by clicking on a picture of a *ship*). For Southern British English

participants, stimuli that were long and had a low F1, i.e. stimuli in the top-right corner of the figure, tended to be classified as /i/, whereas stimuli that were short and had a high F1, i.e. those in the bottom-left corner, tended to be classified as /ɪ/; these listeners can thus be said to rely on both the duration cue and the spectral cue. Scottish English listeners were different, in that they tended to ignore the duration cue: for them, stimuli in the top half of the figure tended to be classified as /i/, those in the bottom half as /ɪ/. Spanish learners of English were again different: some of them acted like the Scots, some like the Southerners, and some showed a pattern not found in either of the native groups, namely ignoring the spectral cue and classifying the stimuli in the right half of the figure as /i/, those in the left half as /ɪ/.

In reality, the results were much more variable than just presented. In order to make sense of the data, we decided to derive from each participant a single quantity, namely the relative extent to which she relied on the two cues.

## 2. Measuring relative cue weighting as an edge difference ratio

The single quantity that E&B used for measuring each participant's relative reliance on duration and spectrum was the slope of a reconstructed *boundary line*, which is a straight line through Figure 1 that separates best the area where the participant predominantly responds /i/ from the area where she predominantly responds /ɪ/. Figure 2 (left) shows how this slope can be computed for an idealized listener, for whom stimuli not on the boundary line are unambiguous.

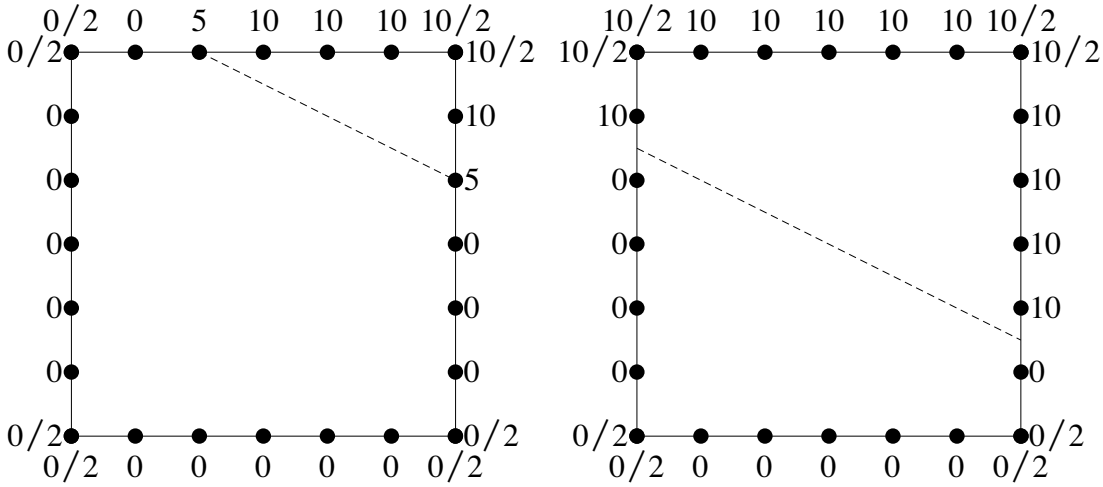


**Fig. 2.** Measuring the slope from the responses on the edges (two idealized listeners).

The left-hand side of Figure 2 shows how a certain idealized listener responds to 10 replications of each of the 24 possible stimuli on the edge. The dashed line shows the boundary line for this listener. Stimuli in the top right are always heard as /i/, hence score 10 /i/ responses and 0 /ɪ/ responses; stimuli to its bottom left are always heard as /ɪ/, hence score 0 /i/ responses (and 10 /ɪ/ responses); the two stimuli that happen to lie on the boundary line score 5 /i/ and 5 /ɪ/ responses.

For the ideal listener in Figure 2 (left), the slope of the boundary line can be computed by two methods. First, it can be computed directly from the line itself. When going 4 units to the right, the line goes down by 2 units, so its slope is (–2)/4 = –0.5. Another way of saying this is that this listener relies on spectrum twice as much

as on duration, so that her *duration/spectrum reliance ratio* is 2/4 = 0.5. But the slope can also be computed indirectly from the responses along the edges. Along the right edge, there are 60 responses in total (of the ten measurements in the top-right corner, five count with the right edge, five with the top edge).[1] Of these, we see 20 /i/ responses. The number of /i/ responses along the left edge is 0. The difference between these values is 20, so that we can compute the listener's *duration reliance* for this rectangle as 20/60 = 33.3%. Similarly, the listener's *spectral reliance* for this rectangle can be computed by subtracting her /i/ scores along the bottom (0) from her /i/ scores along the top (40), which gives a spectral reliance of (40–0)/60 = 66.7%. The ratio of these 'reliances' is 33.3%/66.7% = 0.5, and the slope of the boundary line must therefore be –0.5.

Real listeners do not show the zeroes, fives, and tens of Figure 2, but will have a much more variable pattern, so that their boundary line cannot be drawn by just looking at the numbers around the edges. Therefore, the direct method for computing the slope cannot be used, but an estimate of the slope (and the reliance ratio) can still be computed by the second method, that of dividing the right-left difference by the top-bottom difference, and that is what E&B did for the participants in their experiments.

## 3. The allegedly detrimental 'ceiling effect'

In his critique, Morrison (2005) mentions the word 'ceiling' 19 times. On the right in Figure 2 we can see what he means with this term. The figure shows the boundary line of a second idealized listener. The boundary is in a different location from that of the first listener, but its slope is identical (again –0.5, since six steps to the right is three steps down), and so is therefore its duration/spectrum reliance ratio (0.5). The duration reliance is computed as 45 /i/ responses on the right edge minus 15 /i/ responses on the left edge, divided by 60, gives (45–15)/60 = 50%. The spectral reliance is 60 /i/ responses along the top minus 0 /i/ responses along the bottom, divided by 60, gives 100%. It is this 100% value that Morrison objects to: a spectral reliance can never be greater than 100%, so if one finds a spectral reliance of 100% in a specific case, the value is *at ceiling* and no meaningful measurements can allegedly be based on it. On computing reliance ratios from a spectral reliance that is at ceiling, as in the right hand of Figure 2, Morrison asserts: "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."

But let's nevertheless compute the ratio as in the previous section, i.e. as if the method *were* valid: the reliance ratio is then computed as 50% divided by 100%, which is 0.5, which is the correct value (namely the slope). This perhaps slightly counterintuitive result generalizes to all thinkable cases of boundary lines that intersect the stimulus rectangle: the slope and the reliance ratio can always be correctly estimated as a ratio of response differences, regardless of whether one of the intermediate reliances is at ceiling. It simply works. Morrison's assertion is therefore incorrect.

[1] For E&B, who ignored this corner correction, there were 70 responses along each edge.

But Morrison makes more claims about ceiling effects. He argues that the two 'cue reliances' are ill-defined as observable quantities. This may be correct, and could be used as a criticism against the use of these quantities as experimental end results by Bohn (1995) and Flege, Bohn & Jang (1997). For E&B, however, these quantities are nothing more than intermediate values for computing a single 'reliance ratio', which is the only experimental end result that they discuss. It is just a coincidence that the quantities that E&B needed as intermediate results have the same formal definition as the quantities that Bohn (1995) and Flege et al. (1997) regarded as experimental results. It is simply possible that a ratio of two ill-defined quantities is a well-defined quantity. We return to this subject in §6.

## 4. Measuring reliance from the edge: is logistic regression better?

Now that the 'ceiling effect' has been shown not to exist for the only relevant measure of relative cue weighting, namely the reliance ratio, it becomes interesting to see whether other measures of it than E&B's edge difference ratio could work better.

In order to make plausible that logistic regression works better than taking difference ratios, Morrison gives theoretical arguments and pictures of box plots on E&B's data. But the quality of the two procedures can be assessed directly by measuring their performance on known underlying distributions, and that is what we do in this section.

We generated 500,000 different probability distributions for perceiving /i/, based on bivariate Gaussian distributions for /i/ and /ɪ/. Each of these 500,000 distributions comes with its own *known* reliance ratio.[2] In assessing the accuracy of a method, we divide this underlying reliance ratio into six classes (following E&B's grouping criterion):

| *class* | *duration/spectrum reliance ratio* | *interpretation* |
|---|---|---|
| 1 | more than 4 | exclusively duration |
| 2 | between 2 and 4 | mainly duration |
| 3 | between 1 and 2 | duration and spectrum |
| 4 | between 1/2 and 1 | spectrum and duration |
| 5 | between 1/4 and 1/2 | mainly spectrum |
| 6 | less than 1/4 | exclusively spectrum |

[2] For each of the 500,000 simulations, we started with separate response distributions P(/i/) and P(/ɪ/). For simplicity, we took the duration continuum to run from 1 (left) to 7 (right), and the spectrum continuum from 1 (bottom) to 7 (top). The centres of P(/i/) and P(/ɪ/) were connected by a line with a rising slope that was logarithmically evenly distributed between 1/16 (almost horizontal; listeners almost exclusively use the duration contrast) and 16 (almost vertical; listeners almost exclusively use the spectral contrast). The horizontal as well as the vertical position of the midpoint between the centres of P(/i/) and P(/ɪ/) was uniformly distributed between 2.5 and 5.5. The distance between the centres of P(/i/) and P(/ɪ/) was uniformly distributed between 2.4 and 4.8. The standard deviations of duration ($\sigma_{dur}$) and spectrum ($\sigma_{spec}$) were uniformly distributed between 0.6 and 2.4. The standard deviations were the same for /i/ and for /ɪ/, so that the locations where P(/i/) and P(/ɪ/) are equal form a straight line through the figure (the *equal-likelihood line*). The probability of responding /i/ at a certain point can now be computed as P(/i/) / (P(/i/) + P(/ɪ/)), and the boundary line will fall together with the equal-likelihood line. The known duration/spectrum reliance ratio can be computed as the inverse of the slope of the line that connects the centres of P(/i/) and P(/ɪ/), multiplied by the square of the ratio $\sigma_{spec}/\sigma_{dur}$.

Also known is the probability that each of the 24 stimuli along the edge of the stimulus rectangle is perceived as /i/. We can thus simulate a listener with 240 responses along the edge: each of these 24 probabilities can be used 10 times for computing whether a listener responds /i/ or /ɪ/, so that each of the 24 cells along the edge will end up with a number between 0 and 10. The question now is: how accurate are the two methods in estimating the reliance ratio from the observed response frequencies? For each of the 500,000 simulations, the real underlying class is known from the known reliance ratio. If a method that estimates the reliance ratio from the simulated responses ends up classifying the reliance ratio in a different class, we consider this an error.

The 'edge difference ratio' method used by E&B turned out to classify 87.5 percent of the 500,000 simulated listeners into the correct reliance class. The 12.5 percent errors all consisted of classifying a listener into an adjacent class. The logistic regression method proposed by Morrison, applied to the 24 edge points, turned out to classify 87.9 percent of the listeners correctly.[3] This means that Morrison is correct in stating that the logistic regression method is better, if only by no more than 0.4 percent.

Whereas the two methods under discussion score almost equally well, several other methods turn out to fare worse. The edge difference ratio method on transformed fractions does not work well: for near-logit-transformed fractions, i.e. $\ln((0.01+f)/(1.01-f))$, it scores only 84.7%,[4] and for arcsine-transformed fractions, i.e. $\arcsin(\sqrt{f})$, which equalizes the sizes of the frequentist confidence intervals for binomially distributed data, it scores 86.5%. The worst results are found with linear regression, confirming Morrison's verdict on this method: for non-transformed fractions linear regression scores 82.8%, for near-logit-transformed fractions (which are sometimes used to stand in for true logistic regression!) it scores 82.8% as well, and for arcsine-transformed fractions it scores 83.4%.

The result of the simulations for the edge points is that the edge difference ratio is nearly as good as logistic regression: seeing a reliable difference between the two methods would require doing an experiment with 100,000 participants.

## 5. Taking all data into account, or only part of the data

Another possible problem that Morrison notices is the fact that E&B used only a part of the experimental results. In Figure 1 we can see that every listener had to respond to 37 different stimuli, whereas in Figure 2 we can see that only 24 of these are used for computing the reliance ratio. Therefore, E&B discard approximately one third of the data.[5] We could argue that the 13 discarded stimuli have been used as 'fillers' or 'distractors' in order to prevent the participants from building up a binary durational or spectral opposition on the basis of a bimodal distribution (Maye & Gerken, 2000; Maye, Gerken & Werker 2002), but in reality the stimuli in the middle of the

[3] Logistic regressions were measured with the PRAAT program (Boersma & Weenink 2005).

[4] The 0.01 added to the numerator and denominator is meant to handle fractions of 0 and 1.

[5] Morrison even claims that E&B use only 14 of the 37 stimuli for the measurement of each reliance, but this is irrelevant given that, as argued earlier, the reliance ratio is the only end result, and 24 scores are used to compute it.

rectangle have been used for other purposes, e.g. the categorical perception paradigm (Escudero 2001).

More interesting than the question whether E&B should or should not have taken those 13 stimuli into account, is the more general question of whether the reliance ratio can be computed better on the basis of the scores on the 24 edge points or on the basis of all 49 possible grid points. Figure 3 shows the two set-ups.

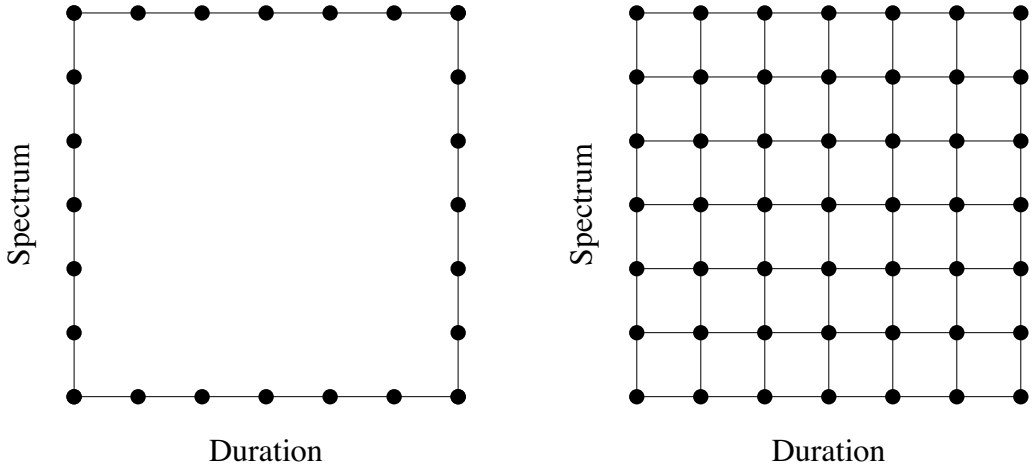


**Fig.3.** Two different stimulus sets: 'edge' (left) and 'grid' (right).

From the previous section we know that the reliance ratio can be classified correctly from the 24 edge points in 87.5% of the cases (edge difference ratio) or 87.9% of the cases (logistic regression). For a test of the measurements from all 49 grid points, we have to distinguish two cases. The first case is that in which the listener responds 10 times to each of the 49 points. The logistic regression method, when applied to 500,000 simulated listeners, classifies the reliance ratios correctly in 90.0 percent of the cases, which is better than either of the measurements based on the edges alone. However, the burden on the participants (490 responses) is more than twice as large as when only the edges are measured (240 responses). To equalize the load on the participants, we consider a second case in which the listener only has to respond 5 times to each stimulus, for a total of 245 responses for the whole grid. In this second case, the logistic regression method classifies the reliance ratio correctly for only 86.1 percent of 500,000 simulated listeners. It thus turns out that if the load for the participants is fixed, it is better to elicit data on the edges (if they do not introduce learning effects) than to elicit data on the whole grid. Table 1 summarizes the results.

**Table 1.** Comparison of various methods for measuring the reliance ratio

| *Method* | *Where* | *Replications* | *Load* | *Quality* |
|---|---|---|---|---|
| Logistic regression | grid | 10 | 490 | 90.0% |
| Logistic regression | edge | 10 | 240 | 87.9% |
| Difference ratio | edge | 10 | 240 | 87.5% |
| Logistic regression | grid | 5 | 245 | 86.1% |
| Linear regression | grid | 10 | 490 | 85.1% |
| Linear regression | edge | 10 | 240 | 82.8% |
| Linear regression | grid | 5 | 245 | 82.2% |

To sum up: linear regression is worse than the other methods (by 4 to 5 percent), measuring the edge alone is better than measuring the whole grid (by 1.8 percent), and logistic regression is better than the difference ratio (by 0.4 percent). As a final note, we can mention that E&B's score would have increased from 87.5% to 89.0% if they had included all the 37 points of Figure 1 in a logistic regression measurement.

## 6. Has two-category assimilation been confirmed?

One of E&B's proposals was that L1-Spanish L2-Scottish English listeners tend to assimilate the English /ɪ/-/i/ contrast to their native /e/-/i/ contrast, hence distinguish Scottish English /ɪ/ and /i/ on the basis of spectral differences alone. This proposal was based on the experimental finding that the duration/spectrum reliance ratios of these listeners were approximately zero: given that the non-beginners identify *ship* and *sheep* correctly without additional training, they must have been able to somehow perceive the contrast, and given that they cannot rely on duration, they must have relied on the spectral cues. Morrison challenges this proposal on the basis of his finding that a comparison of the spectrum-tuned logistic regression coefficients of the various groups shows that these listeners used the spectral cues less than did the native Scottish English listeners, and that the difference in spectral cue use between the L2-Scottish English and the L2-Southern English listeners was not statistically significant.

Morrison's challenge does not go through. First, a failure to reach significance can never deny the existence of a difference. The only fact that remains, then, is that the L2 Scots used spectral cues less reliably than the native Scots. But the spectrum-tuned coefficient ($\beta_{spec}$) by itself does not tell us why, because it will be low if the listener relies more on duration than on spectrum (a steep slope in Figure 1 or 2), and it will be low if the listener's /ɪ/-/i/ boundary is fuzzy. The former possibility is ruled out by the zero duration reliance, so the boundary must be fuzzy.

One should not draw too many conclusions on the basis of the degree of crispness of the /ɪ/-/i/ boundary. A fuzzy boundary can be caused by many factors, including non-perceptual ones, and it seems to be a general property of L2 learners that they perform on a lower level than native speakers on almost any task. This is why E&B, other than Bohn (1995), Flege et al. (1997), and Morrison (2005), stayed away from confounded single-dimensional measures of spectral and durational cue reliance and only considered the reliance *ratio*, a measure based on both dimensions together. When measured by reliance ratios, the difference between the L2-Scottish English and

the L2-Southern English listeners *was* significant, and no other explanation than two-category assimilation is available.

## 7. Grouping

As illustrated above in §4, E&B divided the reliance ratios into six classes. As Morrison notes, the dividing points (4, 2, 1, 1/2, 1/4) between these classes are arbitrary, because the range of the horizontal (duration) axis cannot be equated to the range of the vertical (spectrum) axis. But in fact, E&B's experimental results are likely to be valid only for the stimulus rectangle in Figure 1, because listeners tend to adapt to the size of the rectangle (the 'range effect'; Keating, Mikoś & Ganong 1981). This means that boundary locations and orientations tend to depend on the stimulus set (this is one of the reasons why production and perception data are difficult to compare numerically; see Escudero & Boersma 2003: 83), so that experiment-free division points are out of reach anyway if the stimulus range and the response categories have to be the same for every group of listeners regardless of language background.

Morrison's main criticism on the grouping of reliance ratios into classes is: "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." The specific reason that Morrison mentions is incorrect: the sample size is still 20. Nevertheless, a direct test on the data is indeed generally preferable. However, a parametric test is inappropriate here because the data are not normally distributed. But a non-parametric ranked data test is also inappropriate, because much of the data is at ceiling: although Table 2 (E&B: 561) orders the subjects by reliance ratio, the ordering in the middle regions, where the ratios are between 0.2 and 5, is much more reliable than the ordering at the edges, where we find, for instance, 7 'exclusively duration' participants whose relative ordering must be due to chance. The grouping that E&B performed was one way of making sure that all of these subjects end up effectively unordered. The reason that other possible ways were not considered is that it is plausible that much of the observed data in fact reflects underlying groupings: many listeners will have a reliance ratio that is essentially zero or infinite, because they use a single cue.[6]

## 8. Conclusion

Nearly all of Morrison's (2005) criticisms to Escudero & Boersma (2004) are based on Morrison's implicit assumption that the listening experiment should lead to separate conclusions for durational cue reliance and spectral cue reliance. However, we think that one should be cautious in using such single-dimensional quantities as experimental results, as Bohn (1995) and Flege et al. (1997) did with their edge differences (*temporal effect* and *spectral effect*) and Morrison does with the logistic regression coefficients $\beta_{dur}$ and $\beta_{spec}$. This is because these quantities tend to be confounded with factors that make L2 learners perform worse than native speakers on

[6] Logistic regression would not help here, because it suffers from the same grouping effects.

almost any task. Instead, the ratio of these coefficients (i.e. the edge difference ratio, or $\beta_{dur}/\beta_{spec}$) is a measure of *relative* cue weighting that is intended to maximally cancel out these factors.

Once it is acknowledged that the reliance ratio is the only relevant quantity, the question is how it is to be measured. If prior beliefs about biases in the distributions (such as a belief that reliance ratios cannot be negative) can be ignored, logistic regression is the theoretically optimal method. Given its minute simulated performance lag of 0.4%, E&B's edge difference ratio will also be a correct method for experiments with less than 100,000 participants; its advantage is that it can be computed by hand. However, if one has measured responses on stimuli inside the rectangle, perhaps because one is afraid of bimodal distribution adaptation effects, logistic regression is the only sensible method capable of taking these responses into account. A perhaps surprising result of our simulations, though, is that if one is not afraid of bimodal effects, it turns out to be slightly better to collect *N* responses along the edge than to collect *N* responses distributed over the whole grid.

We thank Morrison for pointing out the virtues of logistic regression, and we will certainly use this method in future analyses, albeit not for preventing a ceiling effect, not for producing single-dimensional reliance measures, but for computing boundary locations and slopes for experiments designed to elicit responses for stimuli on the whole grid, as well as for experiments with natural stimuli, whose spectral and durational properties do not have discrete distributions.

## References


Boersma, P., & Weenink, D. (2005). *Praat: doing phonetics by computer* (Version 4.3.10) [Computer program]. Retrieved April 27, 2005, from http://www.praat.org

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: Theoretical and methodological issues* (pp. 279–304). Timonium, MD: York Press.

Escudero, P. (2001). The role of the input in the development of L1 and L2 sound contrasts: Language-specific cue weighting for vowels. In A. H.-J. Do, L. Dominguez, & A. Johansen (eds.), *Proceedings of the 25th annual Boston University Conference on Language Development* (pp. 250–261). Somerville, MA: Cascadilla Press.

Escudero, P., & Boersma, P. (2003). Modelling the perceptual development of phonological contrasts with Optimality Theory and the Gradual Learning Algorithm. In S. Arunachalam, E. Kaiser, & A. Williams (Eds.), *Proceedings of the 25th annual Penn Linguistics Colloquium*, *Penn Working Papers in Linguistics, 8.1,* 71–85.

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 non-native speakers' production and perception of English vowels. *Journal of Phonetics, 25,* 437–470.

Keating, P., Mikoś, M. J., & Ganong, W. F. III (1981). A cross-language study of range of voice onset time in the perception of initial stop voicing. *Journal of the Acoustical Society of America, 70*, 1261–1271.

Maye, J., & Gerken, L. A. (2000). Learning phoneme categories without minimal pairs. In S.C. Howell, S.A. Fish, & T. Keith-Lucas (Eds.), *Proceedings of the 24th annual Boston University Conference on Language Development* (pp. 522–533). Somerville, MA: Cascadilla Press.

Maye, J., Werker, J. F., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101-B111.

Morrison, G. S. (2005). An appropriate metric for cue weighting in L2 speech perception: Response to Escudero & Boersma (2004). *Studies in Second Language Acquisition* (this issue).