confidence that Spanish learners' perception of Dutch /ɑ/~/a/ is affected by the number of peaks in a training distribution.

3.3. Bayes factors

From having found a *p*-value above 0.05 we cannot draw any conclusions about whether the null hypothesis is true or false. Because we wanted to be able to quantify evidence in favor of both the alternative *and* the null hypothesis, we computed Bayes factors (henceforth "BFs") (e.g., Kass and Raftery, 1995; Rouder et al., 2009; Gallistel, 2009; Kruschke, 2010). A BF denotes the likelihood ratio of the data occurring under the null hypothesis (H_0) versus the data occurring under the alternative hypothesis (H_1):

$$BF_{01} = \frac{p(data|H_0)}{p(data|H_1)}$$

The "01" in this equation refers to H_0 and H_1 respectively. Thus, if $BF_{01} = 10$, the observed data are 10 times more likely to occur if H_0 is true than if H_1 is true; if BF01 = 0.1, the observed data are 10 times more likely to occur if H_1 is true than if H_0 is true. If we assume that H_0 and H_1 are equally likely a priori (as is common and as we do henceforth), the Bayes factor BF_{01} can be said to quantify the evidence in support of H_0 over H_1. Thus, if $BF_{01} = 10$, H_0 is 10 times more likely to be true than H_1 (i.e., the odds are 10 to 1 in favor of H_0); if $BF_{01} = 0.1$, H_1 is 10 times more likely to be true than H_0; (i.e., the odds are 10 to 1 in favor of H_1). Whether a clear choice between the two hypotheses is possible, depends on the magnitude of the Bayes factor. If $BF_{01} > 20$, there is said to be strong support for H_0, and if $BF_{01} < 1/20$, there is said to be strong support for H_1; if, however, BF_{01} lies between 3 and 20, the data are said to moderately favor H_0, and if BF_{01} lies between 1 and 3, the data are said to only trivially favor H_0 (Kass and Raftery, 1995).

In the current paper, the null and alternative hypotheses are defined in terms of the standardized effect size of the difference in the improvement score (= the post-test minus the pre-test accuracy percentage) between the Unimodal and Bimodal groups, i.e., in terms of how much the two groups differ in their improvement of categorization accuracy after as compared to before training. An observed effect size *d* can be calculated as the number of standard deviations difference between two improvement scores:

d = (improvement score of group 1 – improvement score of group 2) / standard deviation

where the standard deviation is the pooled standard deviation.[12] In our case group 1 is the Bimodal group and group 2 the Unimodal group.

The null hypothesis (Figure 5, top) is always the same, namely that there is no difference in the improvement score between the Unimodal and Bimodal groups, and that accordingly the effect size *d* is exactly zero:

H_0: $d = 0$

[12] The pooled standard deviation is calculated as the within-sums-of-squares / (N1+N2-2).

<Insert Figure 5 around here>

The value of the BF depends on the definition of the alternative hypothesis. To accommodate different *a priori* beliefs about the effect size, we computed the BF in four different ways, i.e., with four different alternative hypotheses, which are increasingly less specific about the expected value of the effect size. The first and second alternative hypotheses (H_1 and H_2) include information about the effect size obtained from EBW2011, WER2013 and WB2013; the third and fourth alternative hypotheses (H_3 and H_4) do not. Table 4 provides an overview of the four alternative hypotheses and the resultant BFs, which we will now discuss in detail.[13]

Table 4: The four alternative hypotheses (H) and the resulting Bayes factors (BF).

H	**BF**
H_1: *d* = + 0.50	BF_{01} = 137.86
H_2: *d* is a random value drawn from a uniform distribution between 0 and 1.	BF_{02} = 5.97
H_3: *d* is a random value drawn from a Gaussian distribution with mean 0 and standard deviation 1.	BF_{03} = 5.32
H_4: *d* is a random value drawn from a Cauchy distribution	BF_{04} = 4.73

Alternative hypothesis 1 (Figure 5, second from top) stipulates that the effect size *d* is a specific value:

$$H_1: \quad d = +\,0.50$$

This value of +0.50 is based on effect sizes derived from the improvement scores observed in EBW2011, WER2013 and WB2013, as follows. In EBW2011 and WER2013, one group of listeners was exposed to a non-enhanced bimodal distribution (the Bimodal group), a second group to an enhanced bimodal distribution (the Enhanced group), and a third group to music (the Music group). In WB2013, improvement in categorization was compared between a Music group and two Enhanced groups, one presented with a discontinuous distribution and the other to a continuous distribution. As mentioned in the Introduction (section 1.4), in all three studies the

13 The four Bayes factors can be computed in R (R Core Team, 2013) with the equation **dt** (*t*, *df*) / (**mean** (*weight* * **dt** (*t*, *df*, **ncp** = *d* * **sqrt**(*n*))) / **mean** (*weight*)). In this equation, **dt** is the R function that computes the *t* probability density, and **ncp** is the non-centrality parameter of this density; *t* is the between-groups *t* value of our experiment, i.e. -0.43; *df* is the number of degrees of freedom for a *t* test, i.e. 60+60-2 = 118; *n* is half the geometric mean of the two group sizes (Rouder et al. 2009, p.234), i.e. 60*60/(60+60) = 30; *d* is the hypothesized range of possible effect sizes, and *weight* is the shape of the distribution for all these *d* values. For H_1, *d* is 0.5 and *weight* is 1. For H_2, *d* is (-0.5+1:1e5)/1e5 and *weight* is 1. For H_3, *d* is ((-10e5**width*+0.5):(10e5**width*-0.5))/1e5 and weight is exp(-0.5*(*d*/*width*)^2), where *width* is 1. For H_4, *d* is ((-1000*1e4**width*+0.5):(1000*1e4**width*-0.5))/1e4 and *weight* is 1/(1+(*d*/*width*)^2)), where *width* is sqrt(2)/2 (our equations for H_3 and H_4 are formulated in such a way that they will also work for other values of *width*). At the time of writing the computations for H_3 and H_4 are also available on Rouder's website (http://pcl.missouri.edu/bayesfactor).

improvement score was significantly larger for the Enhanced group than for the Music group. In EBW2011 and WER2013, the improvement score for the Bimodal group was not significantly different from that of the Music group and also not from that of the Enhanced group. For the current analysis, we considered the improvement scores of the previous Enhanced groups as proxies for the expected improvement score of our Bimodal group (which was also exposed to an enhanced bimodal distribution, just as the Enhanced groups in the previous studies; section 1.6). Because it was not clear whether our Unimodal group would behave more similarly to the previous Music groups or to the previous Bimodal groups, we considered the improvement scores of the previous Music and Bimodal groups as proxies for the expected improvement score of our Unimodal group. When calculating the effect sizes observed in the three studies, we used the above-mentioned formula for the effect size *d*, and took a previous Enhanced group as group 1, and either a previous Bimodal group or a previous Music group as group 2. The improvement scores for the Enhanced, Bimodal and Music groups were 6.04% (CI = +2.76 ~ +9.31%), 0.80% (CI = –2.22 ~ +3.83%) and –0.15% (CI = –3.50 ~ +3.21%) respectively in EBW2011, and 6.63% (CI = +4.05 ~ +9.20%), 3.83% (CI = +0.97 ~ 6.68%) and 2.00% (CI = –0.50 ~ +4.50%) respectively in WER2013. The improvement scores for the Enhanced and Music groups in WB2013 were 9.68% (CI=+6.80%~+12.55) and 2.00% (CI= –0.50~+4.50) respectively.[14] The pooled standard deviation for the Enhanced and Bimodal groups was 12.00% in EBW2011 and 9.57% in WER2013. The pooled standard deviation for the Enhanced and Music groups was 12.09% in EBW2011, 8.94% in WER2013 and 9.50% in WB2013. Table 5 shows the resulting effect sizes *d*.

Table 5: Effect size *d* in previous studies (see text).

Previous study	Enhanced–Bimodal	Enhanced–Music
EBW (2011)	+0.44	+0.51
WER (2013)	+0.29	+0.52
WB (2013)		+0.81

The average of the five listed effect sizes is +0.51, which we rounded to +0.50 in hypothesis 1. Notice that this value is explicitly positive, i.e., it reflects the belief that our Bimodal group will have a *higher* improvement score, and thus improve *more* after distributional training than the Unimodal group. The BF calculated on the basis of the null hypothesis versus this first alternative hypothesis expresses strong support for the null:

$BF_{01} = 137.86$

Specifically, BF_{01} indicates that the observed data are 137.86 times more likely to have occurred under H_0 (that *d* is exactly 0), than under H_1 (that *d* is exactly 0.5).

14 The Enhanced group referred to here is the group presented with a continuous enhanced distribution in WB2013 (the Continuous Enhanced group). In WB2013 the group presented with a discontinuous enhanced distribution (the Discontinuous Enhanced group) and the Music group were taken from WER2013.

In alternative hypotheses 2 through 4, the effect size is no longer defined as a specific value, but as a probability density function (Figure 5, as explained below): *d* is expected not to be one specific value, but a random value drawn from a distribution whose form defines the likelihood of that value. In alternative hypothesis 2, the effect size is any value between 0 and 1 with equal probability (Figure 5, middle):

H_2: *d* is a random value drawn from a uniform distribution between 0 and 1.

The hypothesis still includes the information mentioned in Table 5 about previously obtained effect sizes (i.e., all effect sizes in Table 5 fall within the range of the distribution), but it is vaguer about the precise value of the expected effect size than hypothesis 1. Since *d* is defined as 0 or positive, hypothesis 2 expresses the belief that the Bimodal group will improve *at least as much* as the Unimodal group. The BF calculated on the basis of the null hypothesis versus this second alternative hypothesis also expresses support for the null:

$BF_{02} = 5.97$

That is, BF_{02} implies that the observed data are 5.97 times more likely to have occurred under H_0 (that *d* is exactly 0) than under H_2 (that *d* is somewhere between 0 and 1).

Hypotheses 1 and 2 show that previous observations can be incorporated in the alternative hypothesis to different extents, depending on the researcher's belief in the truth value of these observations. Previous observations can also be deemed inappropriate for incorporation in the alternative hypothesis, for example if concerns (such as mentioned in the section 1.2) about the earlier observations create uncertainty about the applicability of the information to the experiment to be performed. In this case, the alternative hypothesis should reflect the assumption that we do not have a clear expectation about the effect size. This is done in alternative hypotheses 3 and 4. In alternative hypothesis 3, the effect size is any value around 0, with values closer to the mean being more likely than values further away from the mean as defined by a Gaussian distribution (Figure 5, fourth from top):

H_3: *d* is a random value drawn from a Gaussian distribution with a mean of 0 and a standard deviation of 1.

Since *d* can be positive, zero or negative, the belief that the Bimodal group will improve at least as much as the Unimodal group, which was inherent in alternative hypotheses 1 and 2, is now dropped. The BF calculated on the basis of the null hypothesis versus the third alternative hypothesis still expresses support for the null:

$BF_{03} = 5.32$

In other words, BF_{03} indicates that the observed data are 5.32 times more likely to have occurred under H_0 (that *d* is exactly 0) than under H_3, (that *d* is a value around zero, whose probability is defined by a Gaussian distribution).

It is possible to be even less specific about the expected value of the effect size than in alternative hypothesis 3, by loosening the belief that the effect size is more likely to occur close to zero. This is done with a Cauchy distribution (for an explanation, see Rouder et al., 2009), as used in alternative hypothesis 4 (Figure 5, bottom):

H_4: *d* is a random value drawn from a Cauchy distribution, with a width of (√2)/2.[15]

Notice in Figure 5 that the tails of the Cauchy distribution are much heavier than those of the Gaussian distribution, thus reflecting a much smaller confidence that the effect size should be relatively close to zero. Again, the BF calculated on the basis of the null hypothesis versus the fourth alternative hypothesis expresses support for the null:

$BF_{04} = 4.73$

Thus, BF_{04} indicates that the observed data are 4.73 times more likely to have occurred under H_0 (that *d* is exactly 0) than under H_4 (that *d* is a value around zero, whose probability is defined by a Cauchy distribution, i.e., with more uncertainty as to the effect size than expressed in the Gaussian distribution used for H_3).

In sum, four different calculations of the Bayes factor, which differ in the extent to which they incorporate *a priori* beliefs about the expected effect size, unanimously support the null hypothesis that there is no difference between bimodally and unimodally trained Spanish participants in improvement of categorization of Dutch [ɑ]- and [a]-tokens. If we follow the interpretation of Bayes factors by Kass and Raftery (1995; section 3.3), the support for the null hypothesis ranges from moderate support (hypotheses 2 through 4, which represent less strong *a priori* beliefs about the effect size than hypothesis 1) to strong support (hypothesis 1, which incorporates the most explicit *a priori* beliefs).

4. Discussion

In the present study we trained Spanish adult participants on a bimodal or a unimodal distribution encompassing the Dutch vowel contrast /ɑ/~/a/, and then tested their improvement in categorization of Dutch [ɑ]- and [a]-tokens after training. For the first time in the research on distributional learning of speech sounds, the bimodal and unimodal distributions had nearly identical dispersions, as defined by the range, standard deviation and edge strength. The results show that Spanish adult participants improve their categorization of Dutch [ɑ]- and [a]-tokens irrespective of the training distribution, and that categorization accuracy does not improve significantly more after exposure to one distribution than after exposure to the other distribution. Additionally, four different Bayes factors (ranging from incorporating *a priori* beliefs about the expected effect size as much as possible to not incorporating previous knowledge at all) provided unanimous evidence for the null hypothesis that there is no difference between bimodally and

---

15 The equation used for the Cauchy distribution is: ((-1000*1e4**width*+0.5):(1000*1e4**width*-0.5))/1e4, where *width* is sqrt(2)/2 (see also note 12).