2-2020

Testing the Perceptual Magnet Effect in Monolinguals and Bilinguals

Michael C. Stern

The Graduate Center, City University of New York

Testing the perceptual magnet effect in monolinguals and bilinguals

by

Michael C. Stern

A master's thesis submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Master of Arts, The City University of New York

2020

Testing the perceptual magnet effect in monolinguals and bilinguals

by

Michael C. Stern

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the thesis requirement for the degree of Master of Arts.

Date	Kyle Gorman Thesis Advisor
Date	Juliette Blevins Acting Executive Officer

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

Testing the perceptual magnet effect in monolinguals and bilinguals

by

Michael C. Stern

Advisor: Kyle Gorman

Previous research has demonstrated an apparent warping of the perceptual space whereby the best exemplars or 'prototypes' of speech sound categories minimize the perceptual distance between themselves and neighboring stimuli in the same category. This phenomenon has been termed the 'perceptual magnet effect' (PME). The present study extends work on the PME to a speech sound category previously unstudied in this paradigm (American English /æ/), and to bilingual speech sound representation and perception. American English monolinguals and Turkish-English bilinguals completed identification tasks, category goodness rating tasks, and same-different discrimination tasks with synthesized vowel sounds from the American English /æ/ category—not present in Turkish—and the Turkish /y/ category—not present in English. Results from the identification and goodness rating tasks provided evidence for internal gradedness within these vowel categories. However, results from the discrimination tasks did *not* provide clear evidence for the PME in either participant group for either set of vowel stimuli. These results, in combination with previous non-replications of the PME, suggest that the PME is not a robust, language-specific effect. However, the cognitive representations of speech sound categories likely have *some* effects on bilingual speech perception, and these categorical effects may play a role in contact-induced sound change.

ACKNOWLEDGMENTS

This thesis is the product of the care and work of many. I want to express my deep gratitude to Kyle Gorman for his invaluable help with project design, data analysis, writing, and encouraging me to develop this project for publication. I want to thank Gita Martohardjono for her advice and support at every stage of this project, for introducing me to linguistic research over two years ago, and for giving me the tools and support to develop myself as a researcher. Thank you to Christen N. Madsen II, my mentor, for helping me frame this project in the context of existing research, helping with experiment design and pilot data analysis, holding me to a high standard, and encouraging me to be the best researcher and person I can be. Thank you to Ilaria Porru and Erjon Xholi for their incredible work recruiting and running participants and preparing data, and to Majd Matar for her translation work. I want to thank those who discussed this thesis with me and offered important guidance during its development, including Juliette Blevins, whose work originally inspired this thesis, Douglas H. Whalen, Jason Bishop, and Kevin D. Roon. Thank you to audiences at SYNC, CIRCL, the 5th Workshop on Sound Change, and the 27th Manchester Phonology Meeting for valuable feedback on earlier stages of this work. Thank you to my colleagues and friends Cass Lowry and LeeAnn Stover for supporting and encouraging me at every step, since we joined the program two and a half years ago. Thank you to my mom, Margaret Hecker, and my dad, Steven Stern, for believing in me as I pursue my sometimes esoteric passion, and to my brother David B. Stern for inspiring me to take my intellectual interests seriously and pursue them through the inevitable setbacks and frustrations. Thank you to Taylor Mascari, whose affection and humor offered a welcome relief from this work when I needed it, and whose caring support gave me the strength to carry this thesis through to completion. Finally, I want to express my appreciation for all those who listened to my weird vowel sounds, whose judgments provided the empirical grounding for this work.

Contents

List of Tables

List of Figures

1 Introduction

Previous experimental work, particularly from Patricia Kuhl and colleagues, has demonstrated an inverse relationship between category goodness ratings and discriminability of auditory stimuli from within the same speech sound category: i.e., the best-rated category exemplars are the most difficult sounds to discriminate from similar sounds in the perceptual space, while poorly rated category exemplars are relatively easier to discriminate from similar sounds once psychophysical distance is controlled for. This finding has been extended to both vowels and liquid consonants (Iverson and Kuhl, 1996) using both behavioral and psychophysiological measures like event-related potentials (Aaltonen et al., 1997). The phenomenon has been termed the 'perceptual magnet effect' (PME), because the best category exemplars act like magnets by perceptually attracting neighboring stimuli (Kuhl, 1991). The PME can also be thought of as a 'warping' of the perceptual space, i.e. a shrinking near category centers and stretching near category boundaries. This is most evident in models of the effect using multidimensional scaling (Iverson and Kuhl, 1995). The PME has been demonstrated in infants as young as 6 months (Grieser and Kuhl, 1989), and has been extended to human adults, but not to rhesus monkeys, suggesting that it is unique to speech perception and not a general auditory effect (Kuhl, 1991). Moreover, it has been shown *not* to manifest in infants learning American English for a speech sound from a foreign language (Swedish /y/), suggesting that it develops as a result of language-specific experience (Kuhl et al., 1992).

The PME has been explained in the framework of prototype theory, which posits that perceptual stimuli are categorized based on relative distances to category 'prototypes', rather than by their positions relative to category 'boundaries' (Mervis and Rosch, 1981; Rosch, 1975, 1978; Samuel, 1982). Research within this framework has demonstrated that prototypical exemplars of cognitive categories 'assimilate' neighboring exemplars; i.e., category exemplars near the prototype tend to be perceived as more similar to the prototype than to perceptually equidistant non-prototypical exemplars. In this way, the PME can be seen as the assimilation of speech sounds to the prototypes of speech sound categories. As an application of prototype theory to speech, the PME has important implications for our understanding of the cognitive representation of speech sounds (i.e., speech sound categories are cognitively represented as prototypes, not as boundaries), the acquisition of these representations in learning (i.e., the learner must induce prototypes, not boundaries), and the real-time perception of speech sounds (i.e., the listener selectively ignores information near prototypes and selectively attends to information near boundaries).

Recent work by Juliette Blevins proposed that the PME may underly contact-induced sound change. Blevins termed this proposal the Areal Sound Pattern Hypothesis (ASPH; Blevins, 2017). The ASPH is an attempt to resolve an apparent 'paradox' of contact-induced sound change, i.e., that it appears to develop via regular internally-motivated processes, but it is triggered by contact with neighboring languages. This paradox is especially apparent in cases where sound patterns diffuse through contact independently of lexical borrowing, so the diffused patterns cannot have spread from borrowed lexemes. According to the ASPH, if speech sound prototypes from one language of a bilingual can act as magnets during perception of their *other* language, this could increase the likelihood of an otherwise language-internal sound change.

However, the PME has not yet been demonstrated in bilinguals, and even in monolinguals, attempts to replicate the PME have had mixed success. For example, there appears to be an enormous amount of individual variability in the results of identification, goodness rating, and discrimination tasks, with some participants actually showing a *positive* relationship between goodness and discriminability, contra the PME (Aaltonen et al., 1997; Sharma and Dorman, 1998). Along these same lines, when stimuli were normed for each participant, no evidence for the PME was found, suggesting that the PME might be an artifact of averaging over participants (Lively and Pisoni, 1997). Moreover, American English monolingual adults have been shown to manifest the PME for a foreign vowel category (Swedish /y/), casting doubt on the language-specific nature of the phenomenon (Frieda et al., 1999). Finally, the majority of evidence for the PME has come from the American English /i/ category, and attempts to extend the findings to other vowel categories have often failed (Sussman and Gekas, 1997; Thyer et al., 2000).

In addition to the doubt cast on the PME from empirical non-replications, the theoretical characterization of the PME has also been criticized on conceptual grounds. It has been proposed that the PME (at least for /i/) might simply be due to general auditory effects, as discriminability has been shown to increase as the first formant (F1) increases, regardless of stimulus categorization (Sussman and Lauckner-Morano, 1995; Macmillan et al., 1988). Since the /i/ exemplars designated as prototypes in studies of the PME had a lower F1 than those designated as non-prototypes (Kuhl, 1991), this general auditory effect could explain the difference in discriminability between prototypes and non-prototypes.

Perhaps the most widespread theoretical criticism of the PME is that it is not different from 'categorical perception' (CP), whereby discriminability is greater across category boundaries than within boundaries (Liberman et al., 1957). While CP and the PME differ theoretically in that CP posits boundaries as the main cognitive representations of speech sound categories, whereas the PME posits prototypes, the only empirical distinction between the theories is that the PME predicts differences in *within-category* discriminability, while CP does not. Therefore, it is crucial that the stimuli used in tests of the PME do not overlap with other categories—which was not ensured in the earliest studies of the PME (Kuhl, 1991). A study that controlled this confound with an identification task failed to replicate the PME (Lively and Pisoni, 1997), although later studies that also used identification tasks to preclude category overlap *were* able to replicate the effect (Iverson and Kuhl, 1995). However, it should be noted that in these identification tasks, participants identified stimuli one at a time, while in the discrimination tasks—the main test of the PME—stimuli were heard in pairs. This is noteworthy because it has been shown that stimuli are identified (i.e., categorized) differently in isolation than in pairs (Thompson and Hollien, 1970). When the stimuli for the identification task were presented in pairs, then the discrimination results could be fully accounted for through CP (Lotto et al., 1998). Still, proponents of the PME have countered that boundary effects (CP) and prototype effects (PME) must arise from different mechanisms, because cross-category discrimination is affected by experimental manipulations which do not affect within-category discrimination (Iverson and Kuhl, 2000).

Further support for the unity of CP and the PME has come from computational modeling. Naomi Feldman and colleagues created a single model that captured both CP and the PME by varying a single parameter τ: the ratio between meaningful variance and random variance in the distribution of sounds in a category, with vowels (for which the PME is most evident) having a high value of τ, and stop consonants (for which CP is most evident) having a low value of τ (Feldman et al., 2009; Kronrod et al., 2016). This computational model has been argued to fit well with other models that capture categorical effects on perception at the algorithmic level (Lacerda, 1995) and the neural implementational level (Geunther and Gjaja, 2000).

Given the so far equivocal results bearing on the PME, more data is needed to shed light on the nature of the phenomenon. Specifically in regard to the ASPH, no data yet exists which addresses the PME in bilinguals, although it has been hypothesized that bilinguals should manifest the PME for both of their languages (Kuhl and Iverson, 1995).

1.1 Overview of the present study

This study extended work on the PME to a speech sound category previously unstudied in this paradigm (American English /æ/), and to bilingual representation and perception. Specifically, we asked the following questions:

1. Will the PME manifest in American English monolinguals for /æ/? This addresses whether the PME is a result of speech sound categorization, or whether it is driven by general auditory perception, and therefore only manifests for high vowels (Sussman and Lauckner-Morano, 1995; Thyer et al., 2000).

2. Will the PME manifest in American English monolinguals for Turkish /y/? This addresses

whether the PME is language-specific, and therefore does not manifest for speech sound categories from a foreign language (Frieda et al., 1999).

3. Will the PME manifest in highly proficient Turkish-English bilinguals for both /y/, from their first-learned language (L1), and /æ/, from their late-acquired second language (L2)? This addresses whether speech sound category prototypes can be acquired later in life (Kuhl and Iverson, 1995), and whether the PME in bilinguals plays a role in contact-induced sound change (Blevins, 2017).

In response to each of these questions, we hypothesize the following:

1. Consistent with Kuhl's Native Language Magnet (NLM) theory of speech sound category acquisition (Kuhl, 1993, 2000; Kuhl et al., 2008), the PME should manifest in American English monolinguals for /æ/, a category from their native language.

2. Consistent with the NLM, the PME should *not* manifest in American English monolinguals for /y/, a category absent from their native language.

3. Consistent with the ASPH (Blevins, 2017) and Kuhl's hypothesis (Kuhl and Iverson, 1995), the PME should manifest in Turkish-English bilinguals for both /y/, from their L1, and /æ/, from their L2.

To address these questions, we conducted identification tasks, category goodness rating tasks, and AX discrimination tasks with synthesized stimuli from the American English /æ/ category and the Turkish /y/ category. A total of six experiments were conducted.

2 Experiment 1: /æ/ identification

The purpose of this experiment was to locate an area of the vowel space that is consistently identified as /æ/ by native speakers of American English in order to ensure that the range of stimuli we used in our tests of the PME did not overlap with any category boundaries, as this could lead to a spurious effect driven by boundary effects on perception (Lotto et al., 1998).

2.1 Method

2.1.1 Participants

Five adult native speakers of American English participated in this experiment. Participants' ages ranged from 18 to 59. All were born and raised in the anglophone US and had no greater than beginner-level proficiency in any language other than English. No participants reported any speech, language, or hearing impairments.

2.1.2 Stimuli

25 vowel stimuli were synthesized using the KlattGrid synthesizer (Klatt and Klatt, 1990) implemented in Praat (Boersma and Weenink, 2019). All of the stimuli had identical durations, pitch contours and root-mean-square (RMS) intensity values based on stimuli from previous studies of the PME (Kuhl, 1991; Iverson and Kuhl, 1995, 2000). Each stimulus was 500 ms in length, and each had a pitch contour rising from 112 Hz to 132 Hz over the first 100 ms, and falling to 92 Hz over the remaining 400 ms. The RMS intensity of all stimuli was normalized to −20 dB in Audacity (Audacity Team, 2019).

The first formant (F1) and second formant (F2) of the stimulus in the center of the matrix (labeled "C3" in Table 1) were set to 660 Hz and 1720 Hz, respectively, based on measurements from the Peterson & Barney database (Peterson and Barney, 1952). F1 and F2 values for the other stimuli were determined using the mel scale (Stevens et al., 1937), argued to approximately equate psychoacoustic distance (Fant, 1973); stimuli were spaced 60 mels apart from each other in F1 and F2. The third formant (F3) for all stimuli was set to 2410 Hz based on the same production study cited above (Peterson and Barney, 1952), but was not varied further; remaining formants were set to the synthesizer's default values.

			1		2		3		4		5
		F1	**F2**	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**
A	**mels**	629	1518	629	1458	629	1398	629	1338	629	1278
	Hz	523	1992	523	1852	523	1720	523	1595	523	1476
B	**mels**	689	1518	689	1458	689	1398	689	1338	689	1278
	Hz	590	1992	590	1852	590	1720	590	1595	590	1476
C	**mels**	749	1518	749	1458	749	1398	749	1338	749	1278
	Hz	660	1992	660	1852	660	1720	660	1595	660	1476
D	**mels**	809	1518	809	1458	809	1398	809	1338	809	1278
	Hz	735	1992	735	1852	735	1720	735	1595	735	1476
E	**mels**	869	1518	869	1458	869	1398	869	1338	869	1278
	Hz	813	1992	813	1852	813	1720	813	1595	813	1476

Table 1: F1 and F2 of each pre-normed /æ/ stimulus in Hz and mels.

The x-axis represents F2, and the y-axis represents F1, by analogy with common two-dimensional representations of the vowel space. For ease of reference, columns are labeled 1-5, and rows are labeled A-E.

2.1.3 Procedure

The experiment was conducted in a quiet room. Instructions and stimuli were presented on a tablet running E-Prime 2.0 (Schneider et al., 2002), which was also used to record participant responses. Stimuli were presented binaurally at a comfortable volume through headphones. At the start of the experiment, participants were told that they would hear some vowels, and that their task was to choose the word that each vowel would fit best in. Each vowel stimulus was presented concurrently with six visual answer choices on the tablet screen: *head*, *had*, *hog*, *hug*, *hood*, and *none of these*. Participants selected their response by touching it on the screen. Participants were able to listen to each stimulus as many times as they liked by selecting *listen again* on the screen. The experiment began with six practice trials, after which the participant was prompted to ask the experimenter any questions they had. Then, each stimulus was presented twice, for a total of 50 trials per participant. Each session lasted roughly 10 minutes.

2.2 Results & Discussion

Table 2 displays the total percentage of identifications as /æ/ (the vowel in *had*) for each stimulus across participants. It is noteworthy that stimulus C3, whose formant values were identical to those recorded in actual productions (Peterson and Barney, 1952), was only identified as /æ/ in 20% of trials, while the more extreme stimuli E1, E2, and E3 were identified as /æ/ in 100% of trials. This is consistent with Keith Johnson's hyperspace effect (Johnson et al., 1993), i.e. that perceptual prototypes tend to be more extreme than average productions.

	1	**2**	**3**	**4**	**5**
A	0	0	20	0	0
B	20	0	0	0	0
C	40	60	20	0	0
D	70	90	80	40	0
E	100	100	100	40	10

Table 2: Percent identification as /æ/ of each stimulus.

3 Experiment 2: /y/ identification

Similar to Experiment 1, the purpose of this experiment was to locate an area of the vowel space that is consistently identified as Turkish /y/ by native speakers of Turkish, in order to avoid confounds arising from boundary effects on perception (Lotto et al., 1998).

3.1 Method

3.1.1 Participants

Five Turkish-English late bilingual adults (ages 18-59) participated in this experiment. Participants were screened with a questionnaire developed in our lab, which collected information pertaining to language use, exposure, and self-reported proficiency. All participants were born and raised in Turkey, learned Turkish as their first language, and arrived to the US at age 17 or later (M = 26.64, SD = 5.40). On a scale from zero (not proficient at all) to six (very proficient), all participants described their proficiency in both Turkish and English as at least four in both speaking and listening (see Table 3). All participants reported no speech, language, or hearing impairments.

	Turkish M (SD)	**English M (SD)**
Speaking	5.97 (0.17)	5.03 (0.77)
Listening	6.00 (0.00)	5.30 (0.68)

Table 3: Self-rated proficiency of Turkish-English bilinguals (out of six).

3.1.2 Stimuli

25 vowel stimuli (see Table 4) were synthesized using the same parameters as the stimuli from Experiment 1. The only difference between these stimuli and those from Experiment 1 were the values of the first three formants. F1 and F2 of the center stimulus C3 were set to 296 Hz and 1635 Hz, respectively; these measurements were derived from a database of productions of /y/ by Turkish speakers (Radisic, 2014). Similar to Experiment 1, step sizes of F1 and F2 between stimuli were 60 mels. F3 for all stimuli was set to 2320 Hz based on the same production study (Radisic, 2014) and was not further varied.

3.1.3 Procedure

The procedure for Experiment 2 was identical to that of Experiment 1 except that the instructions and answer choices were given in Turkish. The five answer choices (in addition to *none of these* and *listen again*) were *kül*, *köl*, *kil*, *kul*, and *kıl*.

3.2 Results & Discussion

Table 5 displays the total percentage of identifications as /y/ (the vowel in *kül*) for each stimulus across participants. The pattern of responses was not as clear as that in Experiment 1, perhaps

		1		2		3		4		5	
		F1	**F2**	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**
A	**mels**	277	1478	277	1418	277	1358	277	1298	277	1238
	Hz	195	1898	195	1763	195	1635	195	1515	195	1400
B	**mels**	337	1478	337	1418	337	1358	337	1298	337	1238
	Hz	244	1898	244	1763	244	1635	244	1515	244	1400
C	**mels**	397	1478	397	1418	397	1358	397	1298	397	1238
	Hz	296	1898	296	1763	296	1635	296	1515	296	1400
D	**mels**	457	1478	457	1418	457	1358	457	1298	457	1238
	Hz	350	1898	350	1763	350	1635	350	1515	350	1400
E	**mels**	517	1478	517	1418	517	1358	517	1298	517	1238
	Hz	408	1898	408	1763	408	1635	408	1515	408	1400

Table 4: F1 and F2 of each pre-normed /y/ stimulus in Hz and mels.

because Turkish /y/ is not a ‘corner vowel’ (its average F2 is substantially lower than that of /i/ (Radisic, 2014)), so the hyperspace effect (Johnson et al., 1993) does not drive a perceptual preference for corner stimuli.

	1	**2**	**3**	**4**	**5**
A	70	50	30	20	10
B	80	90	60	50	0
C	50	90	70	60	10
D	10	20	20	10	0
E	10	0	0	0	0

Table 5: Percent identification as /y/ of each stimulus.

4 Experiment 3: /æ/ goodness rating

The purpose of this experiment was to measure the goodness of each stimulus from the /æ/ matrix, re-synthesized based on the results of Experiment 1, in the category representations of both American English monolinguals and Turkish-English bilinguals. Prototype theory predicts internal gradedness in cognitive categories, so despite all of the stimuli being members of the /æ/ category, some stimuli are predicted to be *better* category exemplars than others. Since both groups of participants had significant exposure to American English, we predicted that both groups would demonstrate similar patterns of subcategorical gradedness.

4.1 Method

4.1.1 Participants

23 American English monolinguals and 20 Turkish-English bilinguals participated in the experiment. Inclusion criteria for the monolinguals were the same as from Experiment 1, and inclusion criteria for the bilinguals were the same as from Experiment 2.

4.1.2 Stimuli

25 vowel stimuli (see Table 6) were synthesized using parameters identical to those from Experiment 1. Only the first and second formants (F1 and F2) differed between these stimuli and those from Experiment 1. This matrix was centered on the midpoint between stimuli E2 and E3 from Experiment 1, which were both identified as /æ/ in 100% of trials. Step sizes of F1 and F2 between stimuli were only 30 mels, rather than 60 mels, in order to ensure that the matrix did not overlap with another vowel category.

			1		2		3		4		5
		F1	**F2**	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**
A	**mels**	809	1488	809	1458	809	1428	809	1398	809	1368
	Hz	735	1921	735	1852	735	1785	735	1720	735	1656
B	**mels**	839	1488	839	1458	839	1428	839	1398	839	1368
	Hz	774	1921	774	1852	774	1785	774	1720	774	1656
C	**mels**	869	1488	869	1458	869	1428	869	1398	869	1368
	Hz	813	1921	813	1852	813	1785	813	1720	813	1656
D	**mels**	899	1488	899	1458	899	1428	899	1398	899	1368
	Hz	854	1921	854	1852	854	1785	854	1720	854	1656
E	**mels**	929	1488	929	1458	929	1428	929	1398	929	1368
	Hz	896	1921	896	1852	896	1785	896	1720	896	1656

Table 6: F1 and F2 of each normed /æ/ stimulus in Hz and mels.

4.1.3 Procedure

Stimuli were presented in a manner identical to Experiments 1 and 2. However, in this experiment, participants were asked to rate how well each stimulus fit in the word *hat* on a six-point Likert scale with endpoints labeled “very good" and “very bad".[1] Similar to Experiments 1 and 2, participants were able to listen to each stimulus as many times as they liked by selecting *listen again*. The experiment began with six practice trials, after which the participant was prompted to ask any questions they had, and then each stimulus was presented twice, for a total of 50 trials. The session lasted roughly 10 minutes.

4.2 Results & Discussion

Goodness ratings were z-scored by participant to account for the different means and variances in each participant's distribution of responses. Then, the mean z-scored goodness rating of each stimulus was calculated by participant group.

4.2.1 American English monolinguals

Monolinguals showed a clear best exemplar at the most peripheral stimulus, E1, and ratings decreased smoothly in both directions moving outward from this stimulus (Table 8). This clear pattern is consistent with the notion of internal gradedness in monolinguals' cognitive representations of the /æ/ category.

[1]We decided to use *hat* in this experiment, rather than *had*, in order to avoid any confounding effects from /æ/ tensing before voiced stops, although this phenomenon has been argued to be disappearing among younger speakers (Coggshall and Becker, 2009). The consistency of this experiment's results with those of Experiment 1 suggests that this difference did not significantly affect participants' responses.

4.2.2 Turkish-English bilinguals

Bilinguals showed basically the same pattern as monolinguals, i.e., goodness ratings increased as F1 and F2 increased (Table 9). This suggests that these highly proficient L2 speakers of English acquired subtle subcategorical information in their representations of the English /æ/ category, during adulthood, closely resembling that represented by English monolinguals. However, bilinguals' pattern in F1 was not as clear as the one seen in monolinguals; e.g., D1 had a higher mean rating than E1, and C1 had a lower mean rating than B1 and A1. This could be due to the less crowded F1 space in the Turkish vowel system,[2] which might cause Turkish speakers to de-emphasize F1 as an informative cue during vowel perception (Iverson et al., 2003; Strange, 2011).

5 Experiment 4: /y/ goodness rating

Similar to Experiment 3, the purpose of this experiment was to measure the category goodness of each stimulus from the /y/ matrix—re-synthesized based on the results of Experiment 2. We predicted that the Turkish-English bilinguals would show evidence for internal gradedness in their representations of this speech sound category from their L1. However, given that the English phonemic inventory does not contain /y/, we did not expect the American English monolinguals to demonstrate a clear pattern of internal gradedness for the stimuli in this category.

5.1 Method

5.1.1 Participants

The same participants from Experiment 3 also participated in Experiment 4. However, three of the monolingual participants were not available to return for this session, so in total, 20 monolinguals

[2]The Turkish vowel system only distinguishes two degrees of phonological height, compared to at least four for American English monophthongs.

and 20 bilinguals participated in this experiment.

5.1.2 Stimuli

25 /y/ stimuli were synthesized (see Table 7), centered on the midpoint between stimuli B2 and C2 from Experiment 2, which were both identified as /y/ in 90% of trials. Step sizes of F1 and F2 between stimuli were 30 mels.

		1		2		3		4		5	
		F1	**F2**	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**	**F1**	**F2**
A	**mels**	307	1478	307	1448	307	1418	307	1388	307	1358
	Hz	219	1898	219	1830	219	1763	219	1699	219	1636
B	**mels**	337	1478	337	1448	337	1418	337	1388	337	1358
	Hz	244	1898	244	1830	244	1763	244	1699	244	1636
C	**mels**	367	1478	367	1448	367	1418	367	1388	367	1358
	Hz	269	1898	269	1830	269	1763	269	1699	296	1636
D	**mels**	397	1478	397	1448	397	1418	397	1388	397	1358
	Hz	296	1898	296	1830	296	1763	296	1699	296	1636
E	**mels**	427	1478	427	1448	427	1418	427	1388	427	1358
	Hz	322	1898	322	1830	322	1763	322	1699	322	1636

Table 7: F1 and F2 of each normed /y/ stimulus in Hz and mels.

5.1.3 Procedure

Given that the English phonemic inventory does not contain /y/, the monolingual participants were asked to rate the fit of each stimulus in the word *coo* (the English /u/ category).[3] For bilingual participants, all instructions and Likert scale labels were in Turkish, and they were asked to rate the fit of each stimulus in the Turkish word *kül*. Otherwise, the procedure was identical to that in Experiment 3.

[3]In identification tasks, /u/ has been shown to be the English category perceptually closest to German /y/ (Polka and Werker, 1994).

5.2 Results & Discussion

The /y/ goodness rating data was analyzed in the same way as the data from Experiment 3.

5.2.1 American English monolinguals

As seen in Table 10, monolinguals tended to give higher ratings to those stimuli with a lower F2, presumably those stimuli closest to a prototypical /u/. However, this pattern was not particularly robust (e.g., stimulus B5 had a relatively low mean rating, despite having a low F2, and B2 had a relatively high rating, despite having a high F2). The lack of a clear pattern is not very surprising from the perspective of prototype theory, given that English monolinguals should not have a category representation for /y/.

5.2.2 Turkish-English bilinguals

Surprisingly, as seen in Table 11, bilinguals did *not* demonstrate a clear pattern of internal gradedness for stimuli in the Turkish /y/ category, a category from their native language. This is reminiscent of the lack of a clear pattern in the results of the identification task (Experiment 2), and could be due to the fact that /y/ is not a peripheral vowel. This possibility is supported by a previous study that failed to demonstrate a clear goodness pattern in representations of the American English /ɪ/ category, another non-peripheral vowel (Sussman and Gekas, 1997).

	1	2	3	4	5
A	−0.39	−0.40	−0.64	−0.94	−0.92
B	−0.02	0.09	−0.21	−0.40	−0.60
C	0.22	0.24	0.19	−0.04	−0.40
D	0.37	0.58	0.60	0.25	−0.18
E	0.91	0.87	0.73	0.48	−0.38

Table 8: Goodness ratings of /æ/ stimuli by American English monolinguals. Darker green shading indicates higher ratings.

	1	2	3	4	5
A	0.26	0.20	−0.39	−0.84	−1.13
B	0.35	0.04	−0.27	−0.48	−0.71
C	0.18	0.36	0.01	−0.30	−0.67
D	0.79	0.49	0.27	−0.03	−0.14
E	0.65	0.65	0.49	0.12	0.07

Table 9: Goodness ratings of /æ/ stimuli by Turkish-English bilinguals.

	1	2	3	4	5
A	−0.25	−0.06	−0.10	−0.18	0.09
B	−0.24	0.22	−0.05	0.04	−0.17
C	0.00	0.04	0.20	0.10	0.34
D	−0.22	0.01	0.11	0.30	0.21
E	−0.31	−0.25	−0.12	0.17	0.14

Table 10: Goodness ratings of /y/ stimuli by American English monolinguals.

	1	2	3	4	5
A	0.08	−0.05	−0.07	−0.18	−0.24
B	−0.06	0.17	−0.05	0.00	−0.33
C	0.10	0.30	0.05	0.31	−0.20
D	−0.20	0.30	0.07	0.17	0.02
E	−0.18	0.05	−0.09	−0.01	0.06

Table 11: Goodness ratings of /y/ stimuli by Turkish-English bilinguals.

6 Experiment 5: /æ/ discrimination

The purpose of this experiment was to measure participants' ability to discriminate each stimulus from adjacent stimuli in the /æ/ matrix. Evidence for the PME would be constituted by two criteria (Iverson and Kuhl, 2000):

1. MINIMUM AT PROTOTYPE: a minimum in discrimination sensitivity at the category prototype, and

2. NEGATIVE CORRELATION: a negative correlation between goodness ratings and discrimination sensitivity.

Given that both groups of participants have cognitive representations of the /æ/ category, we predicted that both groups would demonstrate evidence for the PME.

6.1 Method

6.1.1 Participants

Participants were the same as those from Experiment 3.

6.1.2 Stimuli

Stimuli were the same as those from Experiment 3.

6.1.3 Procedure

The experiment was an AX same-different discrimination task. In each trial, participants listened to two vowel stimuli, separated by a 250 ms silence. Then, they indicated whether the two sounds were exactly the same, or different in any way, by pressing either *f* (same) or *j* (different) on a keyboard attached to the tablet. Each adjacent (in either F1 or F2) pair of stimuli was presented once (80 ‘different’ trials). Each stimulus was also presented with itself for the same number of times that it was presented with a different stimulus (80 ‘same’ trials), for a total of 160 trials. Participants were instructed to select their response as quickly as possible. After reading the instructions, participants completed six practice trials, during which they received feedback on their accuracy and response time on each trial. During the experimental trials, participants received no feedback. Participants were prompted to take a short break after each eighth of the experiment (every 20 trials). The session lasted roughly 20 minutes.

6.2 Results & Discussion

We measured the discriminability of each stimulus using d', a bias-free measure of discrimination sensitivity (Macmillan and Creelman, 1991). d' was first calculated for each stimulus for each participant, and then averaged across participants in each group to create an overall measure of discriminability for each stimulus in each participant group. To further probe the discriminability of each stimulus, we also analyzed response times (RT) on correct responses. RT on AX discrimination tasks has been argued to reflect discriminability, such that longer RTs index decreased discriminability (Pisoni and Tash, 1974; Strange, 2011). Therefore, assuming that discriminability is reflected in RTs as previously proposed, the PME would predict the slowest RT for the prototype, and a positive correlation between goodness ratings and RT. Observations of RT less than 100 ms, as well as those over two times the standard deviation, were removed (Luce, 1991). Then, RT data was log-transformed to address the non-Gaussian distribution (Ratcliff, 1993). Finally, Spearman correlations were used to test the relationship between discriminability and goodness.

6.2.1 American English monolinguals

As seen in Table 12, the stimulus with the lowest d' was A5, which was the stimulus with the second *lowest* goodness rating from Experiment 3, strongly contradicting the MINIMUM AT PROTOTYPE criterion of the PME. To test for the NEGATIVE CORRELATION criterion of the PME, Figure 1A displays the relationship between mean z-scored goodness ratings and mean d' of each stimulus. There was no significant correlation between goodness ratings and d' ($\rho = .21$, $p = .32$). Correlations between goodness ratings and d' were also calculated by participant. After controlling for a false discovery rate of .05 using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), no participant showed a significant correlation. This is inconsistent with the NEGATIVE CORRELATION criterion.

10% of the RT data was removed according to the exclusion criteria described above. As seen in Table 16, the stimulus with the slowest mean RT (C4) was the fourteenth best-rated stimulus from Experiment 3, contradicting the MINIMUM AT PROTOTYPE criterion. As seen in Figure 2A, there was no significant correlation between goodness and log-transformed RT ($\rho = -.23$, $p = .27$). Of the 23 participants, only one showed the predicted positive correlation after the Benjamini-Hochberg adjustment ($\rho = .65$, $p = .02$). None of the other participants showed any significant correlation. Overall, these results are almost entirely inconsistent with either criterion of the PME for monolinguals on /æ/.

6.2.2 Turkish-English bilinguals

Similar to the monolinguals, interestingly, the stimulus with the lowest mean d' among bilinguals was also A5, which was the lowest-rated of all 25 stimuli from Experiment 3 (Table 13). Again, this strongly contradicts the MINIMUM AT PROTOTYPE criterion. As seen in Figure 1B, there was no significant correlation between goodness and d' among bilinguals ($\rho = .09$, $p = .66$). After the Benjamini-Hochberg adjustment, two participants showed positive correlations, although they were not significant (for both, $p = .08$). These positive correlations contradict the NEGATIVE CORRELATION criterion.

15% of the RT data was removed as outliers. As seen in Table 17, again similarly to the monolinguals, the stimulus with the slowest mean RT was C4. This was the nineteenth-rated stimulus from Experiment 3, contradicting the MINIMUM AT PROTOTYPE criterion. As seen in Figure 2B, there was no significant correlation between goodness and log-transformed RT among bilinguals ($\rho = -.17$, $p = .41$), and no individual participant showed a significant correlation after the Benjamini-Hochberg adjustment. Again, this is inconsistent with the NEGATIVE CORRELATION criterion. Overall, the evidence from both the monolinguals and the bilinguals strongly contradicted both criteria of the PME in the American English /æ/ category.

7 Experiment 6: /y/ discrimination

Similar to Experiment 5, the purpose of this experiment was to measure participants' discrimination sensitivity for each stimulus from adjacent stimuli in the /y/ matrix. It has been argued that the PME is a result of language-specific exposure, such that it does not manifest for the sounds of foreign languages (Kuhl et al., 1992). For this reason, since only the Turkish-English bilinguals, and not the American English monolinguals, have had significant exposure to Turkish, we predicted that the bilinguals, but not the monolinguals, would show evidence for the PME in the Turkish /y/ category.

7.1 Method

7.1.1 Participants

Participants were the same as those from Experiment 4.

7.1.2 Stimuli

Stimuli were the same as those from Experiment 4.

7.1.3 Procedure

The procedure was identical to that of Experiment 5, except that for bilinguals, instructions and labels were in Turkish, while all wording was in English for the monolinguals.

7.2 Results & Discussion

The same analyses were performed on these results as those from Experiment 5.

7.2.1 American English monolinguals

As seen in Table 14, the minimum in d' (E5) was the seventh best-rated stimulus from Experiment 4, contradicting the MINIMUM AT PROTOTYPE criterion. As seen in Figure 1C, there was no significant correlation between goodness and d' ($\rho = .18$, $p = .38$), and no participant-level correlation survived the Benjamini-Hochberg correction, inconsistent with the NEGATIVE CORRELATION criterion.

11% of the RT data was removed as outliers. As seen in Table 18, the stimulus with the slowest mean RT (A5) was the tenth best-rated stimulus from Experiment 4, contradicting the MINIMUM AT PROTOTYPE criterion.[4] As seen in Figure 2C, there was a non-significant negative correlation between goodness and RT ($\rho = -.34$, $p = .09$). This correlation is in the opposite direction as that predicted by the PME (recall that higher RTs indicate *decreased* discriminability). However, after the Benjamini-Hochberg adjustment, no individual participants showed this negative correlation. Moreover, one of the 20 participants showed the predicted positive correlation ($\rho = .59$, $p = .04$). Visual inspection of Figure 2C suggested a possible positive correlation between goodness and RT among only the eight best-rated stimuli. A Spearman correlation test confirmed this positive correlation, although it was not significant ($\rho = .69$, $p = .07$). We discuss a possible explanation of this result in the General Discussion section.

[4] A puzzling result should be noted: stimulus E5 had the lowest mean d', indicating low discriminability, but also had the lowest mean log-transformed RT, indicating high discriminability. Perhaps RT on correct responses, compared to d' on all responses, measure different aspects of discriminability.

7.2.2 Turkish-English bilinguals

As seen in Table 15, for bilinguals, the stimulus with the lowest d' (A5) was the second lowest-rated stimulus from Experiment 4, contradicting the MINIMUM AT PROTOTYPE criterion. As seen in Figure 1D, there was no significant correlation between goodness and d' ($\rho = .31$, $p = .13$). However, after the Benjamini-Hochberg adjustment, one of the 20 participants did show the expected negative correlation ($\rho = -.60$, $p = .03$). No other participant showed a significant correlation.

14% of the RT data was removed as outliers. As seen in Table 19, the stimulus with the slowest mean RT (C2) was the third best-rated stimulus from Experiment 4. Although the relatively high rating of C2 could be construed as support for the MINIMUM AT PROTOTYPE criterion, the fact that C2 was not the highest-rated stimulus is, strictly, a contradiction of this criterion. As seen in Figure 2D, there was no significant correlation between goodness and RT ($\rho = -.18$, $p = .39$), and no participant showed a correlation that was significant after the Benjamini-Hochberg adjustment. Overall, neither the monolinguals nor the bilinguals showed strong evidence for the PME in the /y/ category.

	1	2	3	4	5
A	0.37	0.68	0.29	0.24	0.05
B	0.35	0.80	0.42	0.69	0.45
C	0.49	0.91	0.68	0.78	0.38
D	0.51	0.65	0.62	0.51	0.73
E	0.22	0.55	0.42	0.26	0.15

Table 12: *d′* of each /æ/ stimulus among American English monolinguals. Darker blue shading indicates lower *d′*.

	1	2	3	4	5
A	0.42	0.42	0.43	0.18	0.09
B	0.33	0.49	0.44	0.36	0.71
C	0.42	0.50	0.45	0.39	0.61
D	0.49	0.64	0.46	0.58	0.61
E	0.30	0.59	0.20	0.66	0.13

Table 13: *d′* of each /æ/ stimulus among Turkish-English bilinguals.

	1	2	3	4	5
A	0.10	0.56	0.37	0.19	0.05
B	0.15	0.22	0.70	0.56	0.29
C	0.68	0.38	0.22	0.39	0.54
D	0.26	0.64	0.53	0.64	0.78
E	0.30	0.48	0.62	0.26	0.01

Table 14: *d′* of each /y/ stimulus among American English monolinguals.

	1	2	3	4	5
A	0.04	0.15	0.11	0.06	0.00
B	0.16	0.07	0.43	0.23	0.24
C	0.56	0.57	0.33	0.44	0.49
D	0.51	0.44	0.71	0.57	0.27
E	0.47	0.56	0.21	0.12	0.16

Table 15: *d′* of each /y/ stimulus among Turkish-English bilinguals.

	1	2	3	4	5
A	6.02	6.14	6.02	6.17	6.10
B	6.15	6.13	6.01	6.17	6.12
C	6.07	6.12	6.06	6.18	6.11
D	6.03	6.05	6.10	6.11	5.99
E	6.10	5.99	6.06	6.05	5.98

Table 16: Log-transformed RT on each /æ/ stimulus among American English monolinguals. Darker orange shading indicates slower RT.

	1	2	3	4	5
A	5.88	5.86	6.02	5.73	5.89
B	5.83	5.81	5.88	5.83	5.89
C	.85	5.85	5.85	6.03	6.00
D	5.87	5.86	5.89	5.88	5.94
E	5.82	5.96	5.95	5.94	5.80

Table 17: Log-transformed RT on each /æ/ stimulus among Turkish-English bilinguals.

	1	2	3	4	5
A	6.10	6.08	5.98	6.02	6.20
B	6.19	6.01	6.07	6.15	6.14
C	6.12	6.09	5.95	6.13	6.11
D	6.16	6.12	5.97	6.08	5.93
E	5.91	6.14	6.17	5.94	5.85

Table 18: Log-transformed RT on each /y/ stimulus among American English monolinguals.

	1	2	3	4	5
A	5.95	5.86	5.72	6.00	6.01
B	5.70	5.83	5.95	5.97	5.92
C	5.91	6.02	5.87	5.92	5.95
D	5.98	5.86	5.93	5.84	5.88
E	5.90	5.94	5.83	5.97	5.76

Table 19: Log-transformed RT on each /y/ stimulus among Turkish-English bilinguals.

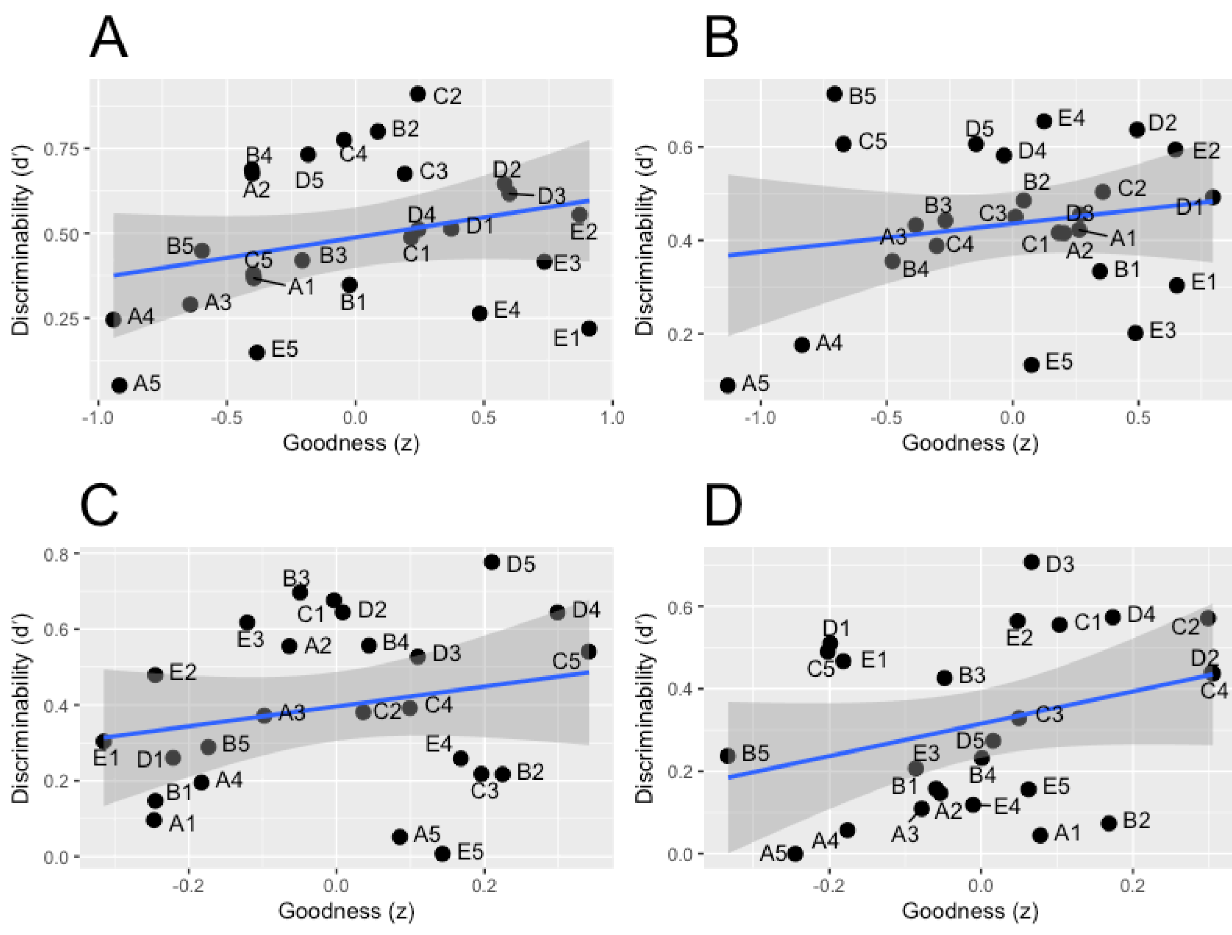

Figure 1: Relationship between goodness and d'.
A: Monolinguals on /æ/. B: Bilinguals on /æ/. C: Monolinguals on /y/. D: Bilinguals on /y/.

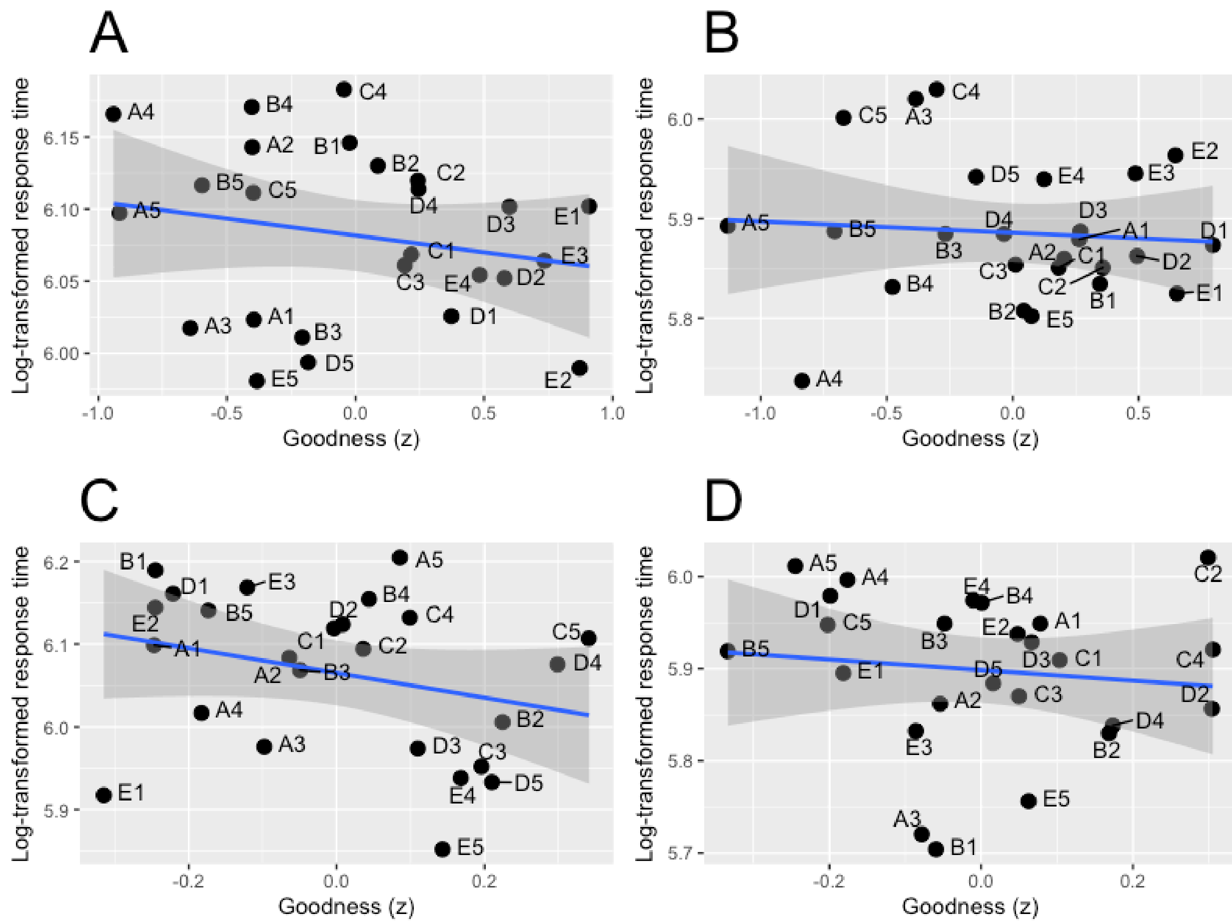

Figure 2: Relationship between goodness and RT.
A: Monolinguals on /æ/. B: Bilinguals on /æ/. C: Monolinguals on /y/. D: Bilinguals on /y/.

8 General Discussion

Both American English monolinguals and Turkish-English bilinguals demonstrated clear patterns of graded category goodness for the /æ/ stimuli. This provides evidence for internal structure in the cognitive representation of this speech sound category, and suggests that listeners consistently attend to subphonemic cues during perception. Moreover, the fact that the bilinguals demonstrated a goodness pattern very similar to that seen in the monolinguals suggests that subphonemic information is acquired during late L2 acquisition, and used in real-time perception of the L2. However, the less clear F1 pattern in the bilinguals compared to the monolinguals suggests that L1 phonemic organization influences the L2 acquisition of subphonemic information, such that information which is more linguistically relevant in the L1 will be more easily acquired in the L2 (Iverson et al., 2003; Strange, 2011).

Surprisingly, the bilinguals did *not* show a clear goodness pattern for the /y/ stimuli. It is likely that this was caused by Turkish /y/ not being a peripheral vowel. Under Keith Johnson's hyperspace effect (Johnson et al., 1993), the best-rated category exemplars tend to be more extreme in the F1–F2 space than actual productions. However, for a non-peripheral vowel like /y/, there is nowhere more extreme in the F1–F2 space that does not encroach on another category. Goodness rating tasks might not be sensitive enough to measure the more specific target ranges of non-peripheral vowel prototypes. This explanation is consistent with the previous lack of a pattern in goodness results for the American English /ɪ/ category (Sussman and Gekas, 1997). This explanation is also consistent with the relatively clear goodness pattern in F2 demonstrated by the monolinguals on the /y/ stimuli. Since the monolinguals were rating the /y/ stimuli as exemplars of the peripheral /u/ category, the hyperspace effect likely drove this preference for stimuli with a low F2.

The results of the discrimination experiments were almost completely inconsistent with the PME for both participant groups in both vowel spaces. In the /æ/ category, neither the monolinguals nor the bilinguals demonstrated any evidence for either criterion of the PME. In the /y/ category, the bilinguals demonstrated a pattern that might be construed as weak support for the MINIMUM AT PROTOTYPE criterion, i.e., the stimulus with the slowest mean RT was quite highly rated, but it was still not the highest-rated stimulus in the category. Interestingly, the only evidence for the NEGATIVE CORRELATION criterion in either vowel category came from the monolinguals on the /y/ stimuli, i.e., among the eight best-rated stimuli, there was a positive correlation between goodness and RT, although it was not significant. This is the exact combination of participant group and vowel space for which the PME was *not* predicted by the NLM, because, by hypothesis, English monolinguals do not have a cognitive representation of /y/. Given that this correlation was only apparent for the eight best-rated stimuli, it could be construed as evidence for the PME in the American English /u/ category. Recall that only the Turkish-English bilinguals participated in the norming identification experiment for /y/, and the American English monolinguals rated the normed stimuli as exemplars of the /u/ category. Therefore, it is possible that, among the monolinguals, only the eight best-rated stimuli were categorized as /u/. Even if we take this as evidence for the PME, however, the puzzle would remain why American English monolinguals would manifest the PME for the /u/ category but not the /æ/ category.

There are a number of methodological differences between our study and earlier studies of the PME (Grieser and Kuhl, 1989; Kuhl, 1991; Kuhl et al., 1992; Iverson and Kuhl, 1995). While we feel these differences are minor, they may be relevant to our largely null results. First, our stimuli covered a smaller range of the perceptual space than those from the original studies. However, this small range was purposely chosen in order to ensure that our stimuli did not overlap with other categories. If this was the reason for the present non-replication, then this would suggest that the effects observed in the original studies were spuriously driven by boundary effects on perception.

Next, the stimulus pairs for the discrimination tasks in the original studies were arranged diagonally (i.e., varying in both F1 and F2), while our stimulus pairs were all arranged horizontally (i.e., varying in either F1 or F2, but never both). Therefore, the Euclidean distances (in the F1–F2 space) between the stimuli in each pair were greater in the earlier studies than in our study. Given that the original studies reported higher values of d' than were observed in our study, it is possible that our discrimination tasks were too difficult for the PME to emerge. However, one study reported evidence for the PME using stimulus pairs that differed in steps of only 15 mels—half the distance between our stimuli (Sussman and Lauckner-Morano, 1995). Therefore, it is unlikely that the present non-replication was caused by the stimulus pairs in the discrimination tasks being too perceptually close.

Finally, we did not address the potential confound caused by stimuli being categorized differently in pairs than in isolation (Lotto et al., 1998). However, this confound would cause spurious evidence *in support* of the PME, that would actually be driven by boundary effects. Therefore, this confound cannot explain the present *absence* of evidence for the PME.

Having addressed potential methodological explanations of our largely null results, we turn now to potential theoretical explanations. First, it is possible that previous evidence for the PME in high vowels has actually just been driven by the general auditory effect that discriminability increases as F1 increases (Macmillan et al., 1988). This explanation is partially supported in our results by a non-significant negative correlation between F1 and log-transformed RT among monolinguals on /æ/ ($\rho = -.39$, $p = .06$; recall that lower RTs indicate higher discriminability), and a significant positive correlation between F1 and d' among bilinguals on /y/ ($\rho = .55$, $p < .01$). However, this explanation is not fully adequate, as neither participant group demonstrated a correlation between F1 and discriminability in their nonnative category, which would be expected of a general auditory effect.

Another potential theoretical explanation of our results relates to the definition of 'prototypes' within the framework of the PME. Although speech sound category prototypes have usually been measured using goodness rating tasks, they have usually been *defined* as the mean of the distribution of sounds in the listeners' input, along some meaningful set of dimensions like mel-scaled F1 and F2 (Feldman et al., 2009; Kronrod et al., 2016; Iverson and Kuhl, 2000). It has been shown that the stimuli identified as prototypes by goodness rating tasks differ quite dramatically from average productions (Johnson et al., 1993), a finding supported by the results of Experiments 1-4 in the present study. Therefore, it is possible that the PME is more accurately conceptualized as applying to the *distributional* prototype, rather than the *best-rated* prototype. We intend to test this hypothesis in future work. Even if it is found that speech sound category prototypes are more accurately identified based on input frequency than goodness ratings, this would still not explain how all of the previous evidence for the PME has come from experiments using goodness ratings.

Taken together, our results suggest that the PME is not a robust, language-specific effect, so it is unlikely to play a role in contact-induced sound change. However, a modified version of the

ASPH (Blevins, 2017) is still plausible, where the PME is replaced with the more robustly attested CP, specifically in its extension to nonnative and L2 perception in Catherine Best's Perceptual Assimilation Model (PAM; Best, 1994; Best and Tyler, 2007). According to the PAM, the speech sound categories of one's native language influence the perception of the sounds of nonnative languages and L2s, such that differences between sounds which fall within the boundaries of a native category will be more difficult to perceive than differences that cross native category boundaries. Although the PAM has so far focused on nonnative and L2 perception, rather than fluent bilingual perception, it provides a clear framework for understanding how the categories of one language can affect the perception of another through the mechanism of CP. In this way, the PAM might provide the ASPH with a more robust mechanism of contact-induced sound change than the PME.

More work is needed to understand the full range of effects of speech sound categories on perception, particularly when the categories are from a different language than the one being perceived. Some steps have already been taken in applying a unified account of categorical perceptual effects incorporating both CP and the PME (Feldman et al., 2009) to nonnative and L2 perception (Barrios, 2013). Research along these lines, particularly in relation to fluent bilingual perception, will likely prove useful in understanding the role of speech perception in contact-induced sound change.

9 Conclusion

We conducted identification, goodness rating, and discrimination experiments with American English monolinguals and Turkish-English bilinguals using synthesized stimuli from the American English /æ/ category and the Turkish /y/ category, varying in subcategorical steps of F1 and F2. We tested the predictions of the NLM theory and the ASPH, respectively: that the PME would generalize to new vowel categories, and to bilingual perception. Our results were almost completely inconsistent with the PME in both vowel categories and participant groups. We conclude, therefore, that the PME is not a generalizable effect, contradicting the NLM theory, and likely does not play a role in contact-induced sound change, contradicting the ASPH. Boundary-based CP effects, applied to cross-language perception through the PAM, likely provide a more robust mechanism of contact-induced sound change. More work will be needed to test this possibility.

References

Aaltonen, O., Eerola, O., Hellström, A., Uusipaikka, E., and Lang, A. H. (1997). Perceptual magnet effect in the light of behavioral and psychophysiological data. *Journal of the Acoustical Society of America*, 101(2):1090–1105.

Audacity Team (2019). Audacity(r): Free audio editor and recorder [Computer application]. Version 2.3.2.

Barrios, S. (2013). *Similarity in L2 phonology*. PhD thesis, University of Maryland.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, 57:289–300.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In Goodman, J. C. and Nusbaum, H. C., editors, *The development of speech perception: The transition from speech sounds to spoken words*, pages 167–224. MIT Press, Cambridge, MA.

Best, C. T. and Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In Munro, M. J. and Bohn, O., editors, *Second language speech learning: The role of language experience in speech perception and production*, pages 13–34. John Benjamins, Amsterdam.

Blevins, J. (2017). Areal sound patterns: From perceptual magnets to stone soup. In Hickey, R., editor, *The handbook of areal linguistics*, pages 88–121. Cambridge University Press, Cambridge.

Boersma, P. and Weenink, D. (2019). Praat: Doing phonetics by computer [Computer program]. Version 6.1.03.

Coggshall, E. L. and Becker, K. (2009). The vowel phonologies of African American and white New York City residents. *Publication of the American Dialect Society*, 94(1):101–128.

Fant, G. (1973). *Speech sounds and features*. The MIT Press, Cambridge, MA.

Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4):752–782.

Frieda, E. M., Walley, A. C., Flege, J. E., and Sloane, M. E. (1999). Adults' perception of native and nonnative vowels: Implications for the perceptual magnet effect. *Perception & Psychophysics*, 61(3):561–577.

Geunther, F. H. and Gjaja, M. N. (2000). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, 100(2):1111–1121.

Grieser, D. and Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology*, 25(4):577–588.

Iverson, P. and Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America*, 97(1):553–562.

Iverson, P. and Kuhl, P. K. (1996). Influences of phonetic identification and category goodness on American listeners' perception of /r/ and /l/. *Journal of the Acoustical Society of America*, 99(2):1130–1140.

Iverson, P. and Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? *Perception & Psychophysics*, 62(4):874–886.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tokhura, Y., Kettermann, A., and Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, 87:B47–B57.

Johnson, K., Flemming, E., and Wright, R. (1993). The hyperspace effect: Phonetic targets are hyperarticulated. *Language*, 69(3):505–528.

Klatt, D. H. and Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2):820–857.

Kronrod, Y., Coppess, E., and Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, 23:1681–1712.

Kuhl, P. K. (1991). Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50(2):93–107.

Kuhl, P. K. (1993). Innate predispositions and the effects of experience in speech perception: The native language magnet theory. In Boysson-Bardies, B. D., de Schonen, S., Jusczyk, P., MacNeilage, P., and Morton, J., editors, *Developmental neurocognition: Speech and face processing in the first year of life*, pages 259–274. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Science*, 97(22):11850–11857.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., and Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society*, 363:979–1000.

Kuhl, P. K. and Iverson, P. (1995). Linguistic experience and the "perceptual magnet effect". In Strange, W., editor, *Cross-language studies of speech perception: A historical review*, pages 121–154. York Press, Baltimore, Maryland.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044):606–608.

Lacerda, F. (1995). The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In Elenius, K. and Branderyd, P., editors, *Proceedings of the XIIIth International Congress of Phonetic Sciences*, pages 140–147.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5):358–368.

Lively, S. E. and Pisoni, D. B. (1997). On prototypes and phonetic categories: A critical assessment of the perceptual magnet effect in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 23(6):1665–1679.

Lotto, A. J., Kluender, K. R., and Holt, L. L. (1998). Depolarizing the perceptual magnet effect. *Journal of the Acoustical Society of America*, 103(6):3648–3655.

Luce, R. D. (1991). *Response times: Their role in inferring elementary mental organization*. Oxford University Press, Oxford, UK.

Macmillan, N. A. and Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge University Press, Cambridge, UK.

Macmillan, N. A., Goldberg, R. F., and Braida, L. D. (1988). Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua. *Journal of the Acoustical Society of America*, 84(4):1262–1280.

Mervis, C. B. and Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, 32:89–115.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2):175–184.

Pisoni, D. B. and Tash, J. (1974). Reaction times to comparisons within and across phoneme boundaries. *Perception & Psychophysics*, 15(2):285–290.

Polka, L. and Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, 20(2):421–435.

Radisic, M. (2014). *An ultrasound and acoustic study of Turkish rounded/unrounded vowel pairs*. PhD thesis, University of Toronto.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114(3):510–532.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology*, 104(3):192–233.

Rosch, E. (1978). Principles of categorization. In Rosch, E. and Lloyd, B. B., editors, *Cognition and categorization*, pages 27–48. Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, 31(4):307–314.

Schneider, W., Eschman, A., and Zuccolotto, A. (2002). E-Prime (Version 2.0). [Computer software and manual]. Psychology Software Tools, Inc.

Sharma, A. and Dorman, M. F. (1998). Exploration of the perceptual magnet effect using the mismatch negativity auditory evoked potential. *Journal of the Acoustical Society of America*, 104(1):511–517.

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8:185–190.

Strange, W. (2011). Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics*, 39:456–466.

Sussman, J. E. and Gekas, B. (1997). Phonetic category structure of [ɪ]: Extent, best exemplars, and organization. *Journal of Speech, Language, and Hearing Research*, 40:1406–1424.

Sussman, J. E. and Lauckner-Morano, V. J. (1995). Further tests of the "perceptual magnet effect" in the perception of [i]: Identification and change/no-change discrimination. *Journal of the Acoustical Society of America*, 97(1):539–552.

Thompson, C. L. and Hollien, H. (1970). Some contextual effects on the perception of synthetic vowels. *Language and Speech*, 13(1):1–13.

Thyer, N., Hickson, L., and Dodd, B. (2000). The perceptual magnet effect in Australian English vowels. *Perception & Psychophysics*, 62(1):1–20.