Supplementary Materials

SECTION 1 PARTICIPANT INFORMATION

Table S1

Participants' Linguistic Information

	English (1)	**Spanish-English (2)**	**Chinese-English (3)**	**Difference**	**Effect size (Cohen's d)**
English	**English**	**English**	**English**		
Listening	9.70 (1.40)	9.57 (0.85)	6.79 (1.76)	1 vs. 3 ** 2 vs. 3 **	1.87 2.15
Speaking	9.61 (1.53)	9.30 (1.25)	6.55 (1.87)	1 vs. 3 ** 2 vs. 3 **	1.82 1.80
Reading	9.71 (1.38)	9.34 (1.03)	7.08 (1.63)	1 vs. 3 ** 2 vs. 3 **	1.76 1.73
AoA (mean)	0.22 (1.15)	3.55 (3.72)	7.39 (3.93)	1 vs. 2 ** 1 vs. 3 ** 2 vs. 3 **	1.21 2.70 1.01
AoA (range)	0-7 (7)	0-18 (18)	0-17 (17)		
Non-English Language (Spanish or Chinese)		**Spanish**	**Chinese**		
Listening	-	9.00 (1.48)	9.75 (0.81)	**	0.60
Speaking	-	7.99 (1.87)	9.56 (1.03)	**	0.98
Reading	-	7.66 (2.02)	9.81 (0.58)	**	1.32
AoA (mean)	-	2.15 (4.67)	0.69 (2.30)	n.s.	0.37
AoA (range)	-	0-21 (21)	0-13 (13)		
Later-acquired Language (English, Spanish, or Chinese)		**English or Spanish**	**English or Chinese**		

Listening	-	9.05 (1.42)	6.74 (1.69)	**	1.50
Speaking	-	8.43 (1.95)	6.42 (1.80)	**	1.06
Reading	-	8.54 (1.77)	6.95 (1.49)	**	0.96
AoA (mean)	-	5.39 (4.72)	7.40 (3.93)	*	0.45
AoA (range)	-	0-21 (21)	0-17 (17)		
Total number	55 (45 F)	56 (42 F, 1 Non-Binary)	38 (24 F)		

Note. For Listening, Reading, and Speaking, numerical values depict participants' mean self-rated proficiency in the corresponding domains out of a scale from 1 to 10 (standard deviations in the parentheses). AoA represents the age of acquisition. The Non-English Language category stands for Spanish for Spanish-English bilinguals and Mandarin-Chinese for Chinese-English bilinguals. The Later-acquired Language category refers to the language that is acquired/learned later (Spanish or English for Spanish-English bilinguals, and Chinese or English for Chinese-English bilinguals). Significant between-group differences are denoted as ** ($p < .01$), * ($p < .05$), or n.s. (not significant).

Table S2

Participants' Demographic Information

	English	Spanish-English	Chinese-English
Age			
mean (SD)	19.20 (1.70)	20.14 (3.52)	22.95 (5.09)
range	17-29 (12)	18-36 (18)	18-37 (19)
Race/Ethnicity (%)			
White	39 (70.9)	6 (10.7)	0 (0.0)
Black and African American	4 (7.3)	0 (0.0)	0 (0.0)
Asian	5 (9.1)	1 (1.8)	38 (100.0)
Hispanic or Latino	3 (5.5)	44 (78.6)	0 (0.0)
Multi	3 (5.5)	5 (8.9)	0 (0.0)
Other	1 (1.8)	0 (0.0)	0 (0.0)
Self-education (%)			
High School	13 (23.6)	18 (32.1)	14 (36.8)
Some College or Associates	41 (74.6)	33 (58.9)	8 (21.1)
Bachelors	1 (1.8)	5 (8.9)	5 (13.2)
Advanced education (Masters or PhD)	0 (0.0)	0 (0.0)	11 (29.0)
Parental Education (%)			
High School	4 (7.3)	34 (60.7)	10 (26.3)
Some College or Associates	12 (21.8)	7 (12.5)	7 (18.4)

Bachelors	17 (30.9)	11 (19.6)	17 (44.7)
Advanced education (Masters or PhD)	22 (40.0)	4 (7.1)	4 (10.5)
Total number	55 (45F)	56 (42 F, 1 Non-binary)	38 (24F)

Note. For Age-mean, each cell represents the mean age of acquisition per language group (standard deviation in parenthesis). For Age-range, each cell denotes the minimum and the maximum number of age per language group (range = max - min in parenthesis). For Race/Ethnicity, Self-education, and Parental education (the highest education of both parents), each cell represents the number of participants per language group (proportion in parenthesis).

SECTION 2 WORD SET AND TONAL INFORMATION

Table S3

Novel Word Sets

Words	No.	Set 1	Set 2
Word 1 (1-8)	1	duti	muda
	2	kami	kuti
	3	bimu	maku
	4	miga	gadu
	5	gadi	gubi
	6	tida	dita
	7	kadu	dumi
	8	tubi	tiga
Word 2 (9-16)	9	batu	buka
	10	dabu	damu
	11	kudi	kadi
	12	bita	kagu
	13	magu	migu
	14	gumi	gami
	15	mika	bitu
	16	tika	tibu

Note. In each set, words from 1 to 8 (or 9 to 16) represent the word pool of one word for a referent. Each word is randomly selected from each pool to form a pair (e.g. No. 1 "*duti*" and No. 15 "*mika*" in Set 1 can be paired with the same referent). Only Word 2 in each set is embedded with tonal contours (T2-T4 or T4-T2) in the Cued condition.

Table S4

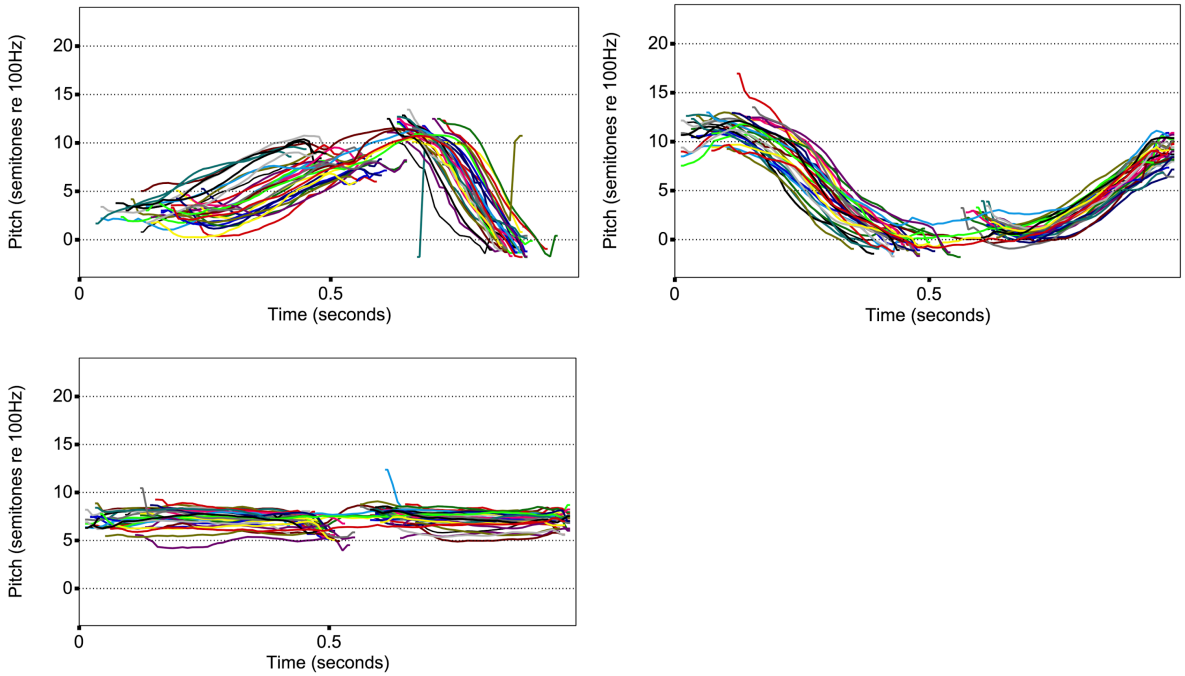
Novel Word Composition By Syllabic Position

	Set 1							**Set 2**						
	Word 1 (1 to 8)		**Word 2 (9 to 16)**		**Total** (Word 1 & Word 2)			**Word 1 (1 to 8)**		**Word 2 (9 to 16)**		**Total** (Word 1 & Word 2)		
Syllable	initial	final	initial	final	initial	final	**initial & final**	initial	final	initial	final	initial	final	**initial & final**
ba	0	0	1	0	1	0	**1**	0	0	0	0	0	0	**0**
bi	1	1	1	0	2	1	**3**	0	1	1	0	1	1	**2**
bu	0	0	0	1	0	1	**1**	0	0	1	1	1	1	**2**
ka	2	0	0	2	2	2	**4**	0	0	2	1	2	1	**3**
ku	0	0	1	0	1	0	**1**	1	1	0	0	1	1	**2**
ma	0	0	1	0	1	0	**1**	1	0	0	0	1	0	**1**
mi	1	1	1	1	2	2	**4**	0	1	1	1	1	2	**3**
mu	0	1	0	0	0	1	**1**	1	0	0	1	1	1	**2**
ga	1	1	0	0	1	1	**2**	1	1	1	0	2	1	**3**
gu	0	0	1	1	1	1	**2**	1	0	0	2	1	2	**3**
da	0	1	1	0	1	1	**2**	0	1	1	0	1	1	**2**
di	0	1	0	1	0	2	**2**	1	0	0	1	1	1	**2**
du	1	1	0	0	1	1	**2**	1	1	0	0	1	1	**2**
ta	0	0	0	1	0	1	**1**	0	1	0	0	0	1	**1**
ti	1	1	1	0	2	1	**3**	1	1	1	0	2	1	**3**

Note. The tables below demonstrate the novel word composition for the two novel word sets for the CSWL tasks. The composition is based on the position of a single syllable in a bisyllabic word ("***bu***" in syllable initial position, e.g. "***bu****ka*", or in syllable final position, e.g. "*ti****bu***"). All syllables are present in all tested languages. /ki/ and /gi/ are expelled due to their non-existence in Mandarin Chinese phonotactics. Each consonant-vowel combination (e.g. "*bu*") appeared approximately the same number of times in word initial position (e.g. "***bu****ka*") and in word final position (e.g. "*ti****bu***") in each word set.

Figure S1

Depiction of Pitch Variation for the Recorded Words in Three Tonal Formats



Note. **Top-left panel:** Pitch variation for words embedded with Mandarin Tone 2-4 contour (rising-falling). **Top-right panel:** Pitch variation for words embedded with Mandarin Tone 4-2 contour (falling-rising). **Bottom-left panel:** Pitch variation for words with no Mandarin tonal contours. Each line denotes an individual novel word.

Table S5

Acoustic Properties for the Recorded Words in the Three Tonal Formats

Tonal pattern	Condition	Duration (s)	Pitch (Hz)	Intensity (dB)
Mandarin Tone2-4 (rising-falling)	Cued	0.99	140.89	62.00
Mandarin Tone4-2 (falling-rising)	Cued	0.99	135.54	65.31
Non-tone	Cued and Uncued	0.99	153.14	66.39
	Total	0.99	142.38	64.53

Note. The table shows the detailed acoustic properties (duration, pitch, and intensity) of the recorded words in the three tonal formats (Tone 2-4, Tone 4-2, and non-tone).

SECTION 3 JUDGMENT OF TONAL STIMULI

Below lists the experimental design of judging the tonality of the word stimuli from a group of naïve listeners who were unfamiliar with Mandarin lexical tones ($N = 65$), as well as the data analysis and results. Results suggested that listeners unfamiliar with the Mandarin tones were able to judge the words embedded with Mandarin tones as different from those without.

In order to test whether the two types of word stimuli (tonal and non-tonal) in the study were perceived as words stemming from two language sources, another group of naïve listeners who were not familiar with Mandarin tones ($N = 65$) and who had not participated in the current study were recruited online (English monolinguals $n = 32$, Other-English bilinguals $n = 33$). In each trial, three distinctive bisyllabic words from the word set inventory (see Supplementary Materials **Table S3** for novel word sets) were auditorily displayed in sequence. The three words were either similar in containing no tones (*control* trials), or one word differed from the other two regarding whether a lexical tonal contour was present or not (*test* trials).

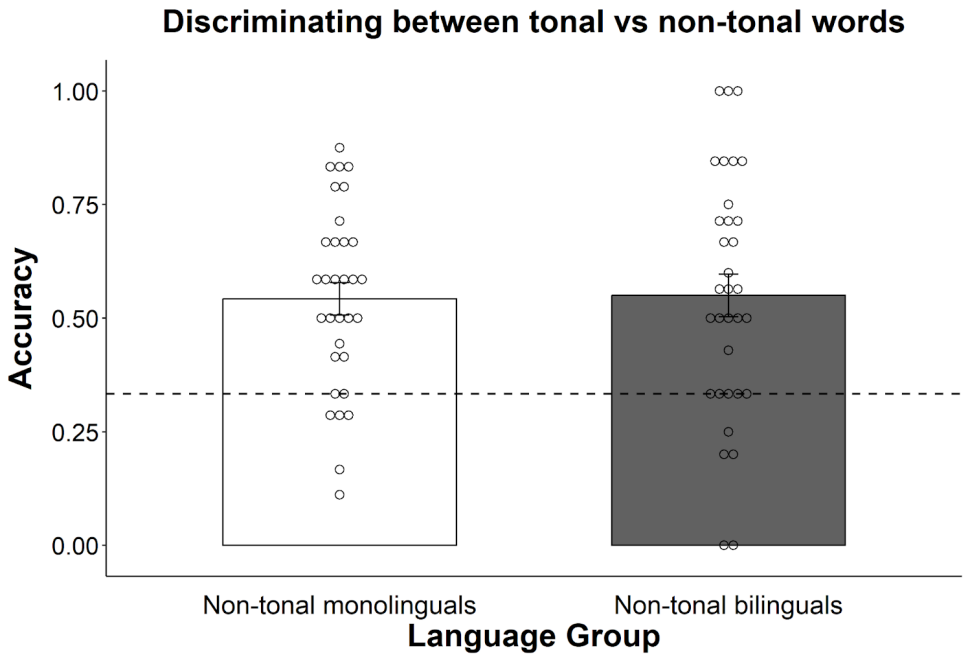
Instructions in each trial went as "*Two of the three words belong to the same language. Find the one that does not belong*"; participants were asked to click onto the one that they subjectively judged as from a different language. In the *test* trials, one of the three words was embedded with a lexical tonal contour (Mandarin Tone 2-4 or Mandarin Tone 4-2) while the other two were not; or one of the three words was not embedded with a tonal contour while the other two were embedded with the same tonal contours. An example of *test* trials could be "*duti*", "*kami*", and "*bàtú*"* (Tone 4-2); another example of *test* trials could be "*duti*"*, "*kámi*" (Tone 2-4), and "*bátù*" (Tone 2-4). The asterisk in each example indicated the target word in the trial. The target words had an equal chance as the firstly, secondly, or lastly presented word. In the *control* trials, all three words were not embedded with any lexical tonal contours. An

example of a *control* trial could be "*duti*", "*kami*", and "*batu*". Since none of the words were different from the others (in lexical tonal contour), the target word choice was expected to be random. Each participant judged 40 *test* trials and 10 *control* trials.

First, we analyzed the target selection in the *test* trials, in which participants successfully chose the target. We conducted a one-sample t-test to compare the target-selection rate against chance (⅓). Results showed that target-selection was significantly above chance (M = .57, SD = .22, $t(64) = 8.73$, $p < .001$). Also, English monolinguals (M = .56, SD = .18) and Other-English bilinguals (M = .58, SD = .25) did not differ from each other in selecting the target ($F(1,63)$ = .14, p = .710). **Figure S2** depicts the results for the *test* trials. Second, we analyzed the target-selection rate in the *control* trials where the three words were all non-tonal. Since there was no target in the *control* trials (expected chance = 0), we instead analyzed participants' selection in word position (at the first, the second, or the third position) in a one-way ANOVA. Results showed that at a random selection, participants selected word position differently ($F(2,110) = 5.65$, p = .005, η^2 = .09). Specifically, when with no target, the lastly presented word (M = .41, SD = .16) were more likely to be chosen compared with the firstly presented (M = .31, SD = .12, p = .001) or the secondly presented word (M = .33, SD = .18, p = .005), with Bonferroni adjusted p-value. In all, the results corroborated the premise that naïve listeners can utilize the presence of lexical tones as an effective cue to mark off different languages. It serves as a strong support that the two words differed in Mandarin lexical tones, in the *Cued* condition of the current study, can simulate two language inputs as that from a bilingual environment.

Figure S2

Rating of Stimuli from Naïve Listeners



Note. The figure depicts, in a rating of stimuli study, the mean accuracy (and standard error of the mean) as a function of the Language Group (Non-tonal English monolinguals and Non-tonal English-Other bilinguals) when discriminating between the tonal vs non-tonal words (the word stimuli used in the main study). Above chance performance (chance = ⅓ denoted as the dashed line) represents a success in picking the one word with (or without) Mandarin lexical tonal contours among the other two words without (or with) Mandarin tonal contours. Dots represent individual participants' data points.

SECTION 4 RANDOM PRESENTATION

Table S6 lists one example of the randomized presentation of Word 1 (W1) and Word 2 (W2) for each given object during the training. The design intends to minimize and reduce the order effects of encountering the first or second presented word to an object, as the order effect would have been washed out throughout the training.

Table S6

The Randomized Order to Present Word 1 and Word 2 for Each Object During the Training (Example of Test List1)

Test list 1								
Order of word-object	**Object 1**	**Object 2**	**Object 3**	**Object 4**	**Object 5**	**Object 6**	**Object 7**	**Object 8**
1st time	**Word 2**	Word 1	Word 2	Word 1	Word 1	Word 2	Word 2	Word 1
2nd time	**Word 1**	Word 2	Word 1	Word 1	Word 2	Word 2	Word 1	Word 2
3rd time	Word 2	Word 1	Word 2	Word 1	Word 2	Word 1	Word 1	Word 1
4th time	Word 1	Word 1	Word 2	Word 1	Word 1	Word 2	Word 2	Word 1
5th time	Word 2	Word 2	Word 1	Word 1	Word 2	Word 1	Word 1	Word 2
6th time	Word 2	Word 2	Word 1	Word 2	Word 2	Word 2	Word 2	Word 1
7th time	Word 1	Word 2	Word 2	Word 2	Word 1	Word 1	Word 2	Word 1
8th time	Word 2	Word 1	Word 1	Word 2	Word 1	Word 1	Word 2	Word 2
9th time	Word 2	Word 2	Word 2	Word 1	Word 1	Word 1	Word 1	Word 2
10th time	Word 1	Word 1	Word 1	Word 2	Word 1	Word 1	Word 1	Word 1
11th time	Word 1	Word 2	Word 2	Word 2	Word 2	Word 2	Word 2	Word 2
12th time	Word 1	Word 1	Word 1	Word 2	Word 2	Word 2	Word 1	Word 2

Note. Each object co-occurs most frequently with two words (Word 1 and Word 2), each for 6 times and with a total of 12 times. The table lists the example of one test list we used in the study (the total number of test lists were 8). The columns list all objects trained during the training (a total number of 8); the rows represent the order where each object co-occurs with one of its two words, Word 1 or Word 2. For instance, the bold **Word 1** and **Word 2** in the table means that for Object 1, the first time it co-occurs with Word 2; the second time it co-occurs with Word 1, etc.

SECTION 5 OTHER DATA ANALYSIS

1 Pre-registered Analysis Plan and Results

The section below provides the results generated according to the pre-registered data analysis plan, such as t-tests and ANOVA (see pre-registration in OSF: https://osf.io/bv5ts). All R scripts and datasets for the pre-registered results are openly accessible (OSF: https://osf.io/kq72m/). The results from the main text by using GLMM are consistent with these from the pre-registered data analysis by using t-tests and ANOVA, except for one difference in the main effect of Group (see below). We opted for consistently reporting GLMM results as a more conservative approach.

1.1 T-tests for Word Learning Against Chance

According to the pre-registered analysis plan, we first examined if adults were successful at learning. For each participant, we aggregated the scores across test trials (each test trial was scored either 0-incorrect or 1-correct) to form four mean accuracy rates for each word type (W1 and W2) and each condition (Cued and Uncued). We then conducted one-sample t-tests separately by language group to compare accuracy rates to chance (0.25). Each language group demonstrated successful learning across word types and conditions (see **Table S7**). Further, aggregating the accuracy rates across word types, conditions, and groups, learners on average were above chance in learning (M = .38, SD = .20, $t(148) = 14.13$, $p < .001$, $d = 1.16$).

Table S7

Mean Accuracy (SD and Effect Size) for Word Type, Condition, and Language Group

	Uncued Condition						**Cued Condition**					
	W1			**W2**			**W1**			**W2**		
	Acc (SD)	***t***	***d***	**Acc (SD)**	***t***	***d***	**Acc (SD)**	***t***	***d***	**Acc (SD)**	***t***	***d***

English Monolingual	.37 (.24)	3.65***	.49	.38 (.18)	5.10***	.69	.37 (.24)	2.99**		.36 (.16)	5.12***	.69
Spanish-English Bilingual	.41 (.18)	6.90***	.92	.38 (.17)	5.99***	.80	.37 (.19)	4.59**	.61	.36 (.24)	3.55***	.47
Chinese-English Bilingual	.38 (.20)	4.25***	.69	.38 (.17)	4.95***	.80	.42 (.17)	6.48***	1.05	.47 (.23)	6.12***	.99

Note. The table depicts accuracy rates of learning (and standard deviations parentheses) by word type, condition, and language group; statistics of comparing the accuracy against chance (.25) are shown by t-values and effect sizes (cohen's d) (*p < .05, **p < .01, ***p < .001). W2-Tone in the Cued condition was embedded with Mandarin lexical tonal contours, while the other words were non-tonal.

1.2 Three-Way ANOVA for Word Learning as a Function of Word Type, Condition, and Group

According to the pre-registered analysis plan, we conducted a three-way mixed ANOVA to measure the impact of Word Type (W1 and W2), Condition (Uncued and Cued), and Group (English monolingual, Spanish-English bilingual, and Chinese-English bilingual) on word accuracy rate. The main effect of Group was significant (F(2, 146) = 3.07, p = .049, η^2 = .01). The main effects of Word type (F(1, 146) = .38, p = .539, η^2 = .00) and Condition (F(1, 146) = .01, p = .912, η^2 = .00) were not significant. There was a significant Condition×Group interaction (F(2, 146) = 3.69, p = .027, η^2 = .01). All other interactions were not significant: Word type×Condition (F(2, 146) = 1.20, p = .275, η^2 = .00), Word type×Group (F(2, 146) = .88, p = .419, η^2 = .00), and Word type×Condition×Group (F(2, 146) = .07, p = .937, η^2 = .00). See **Figure S3**.

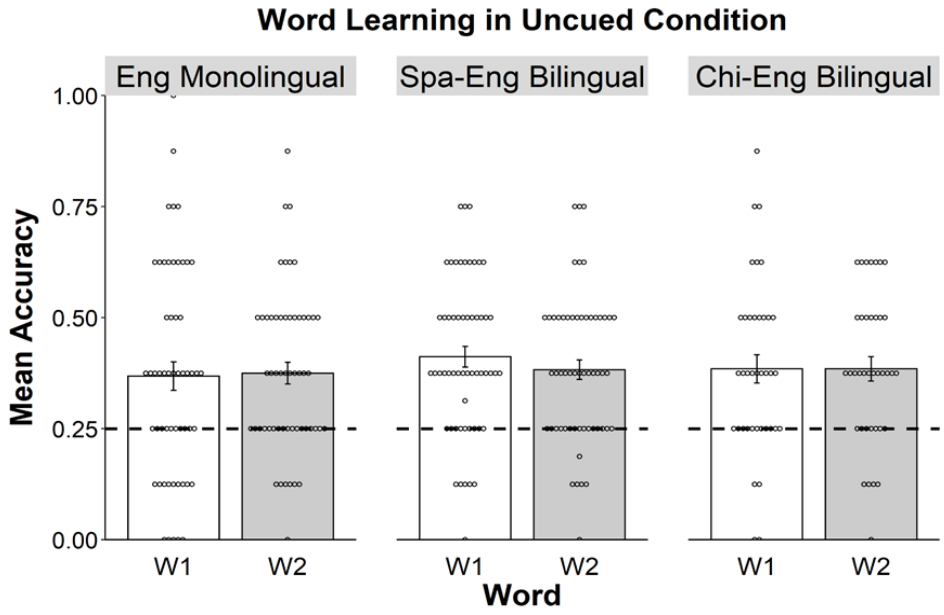
To better understand the interaction between language group and condition, we conducted post-hoc comparisons between the language groups at the level of each condition, with p-value Bonferroni adjustments. Results showed that in the Cued condition, the three language groups

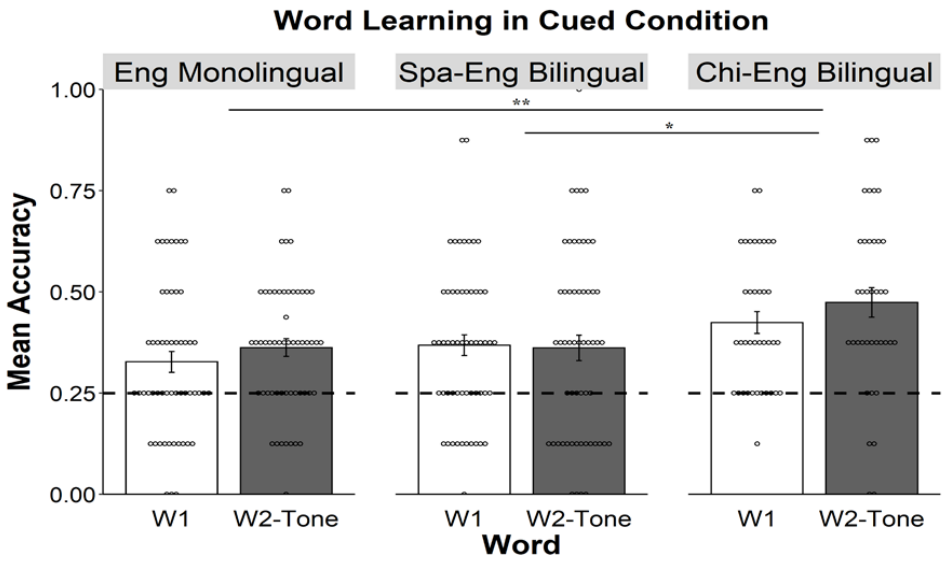
significantly differed ($F(2, 146) = 6.61$, $p = .002$, $\eta^2 = .08$). Specifically, pairwise t-tests showed that Chinese-English bilinguals ($M = .45$, $SD = .20$) were significantly higher in accuracy rate than Spanish-English bilinguals ($M = .37$, $SD = .21$; $p_{adj} = .015$) and than English monolinguals ($M = .35$, $SD = .18$; $p_{adj} = .002$), while Spanish-English bilinguals did not differ from English monolinguals ($p_{adj} = 1.000$). There were no differences in accuracy among groups in the Uncued condition [English monolinguals ($M = .37$, $SD = .21$), Spanish-English bilinguals ($M = .40$, $SD = .17$), Chinese-English bilinguals ($M = .39$, $SD = .18$), $F(2, 146) = .41$, $p_{adj} = 1.000$, $\eta^2 = .01$].

Of notice that the pre-registered plan also showed a significant main effect of Group, which was inconsistent of the results reported in the main text using GLMM (the main effect of Group was marginally significant). We kept it consistent by reporting the results of GLMM as the final data, as GLMM serves as a more conservative approach by considering by-subject and/or by-item level random effects.

Figure S3

Mean Accuracy in the Uncued and Cued Conditions by Word Type and Group.





Note. Mean accuracy (and standard error of the mean) for W1 and W2 in the Uncued condition (**upper pane**l) and Cued condition (**lower panel**) for three language groups. Dashed line denotes chance performance (0.25). Asterisks denote significant between-group differences (*$p < .05$, **$p < .01$), Bonferoni adjusted. Only W2 in the Cued condition (in dark grey) was embedded with Mandarin lexical tones (e.g. "*tíkà*"), while the others were non-tonal (in white or gray) (e.g. "*batu*"). Dots stand for individual data points.

1.3 T-tests for Learning One or Two Labels Against Chance

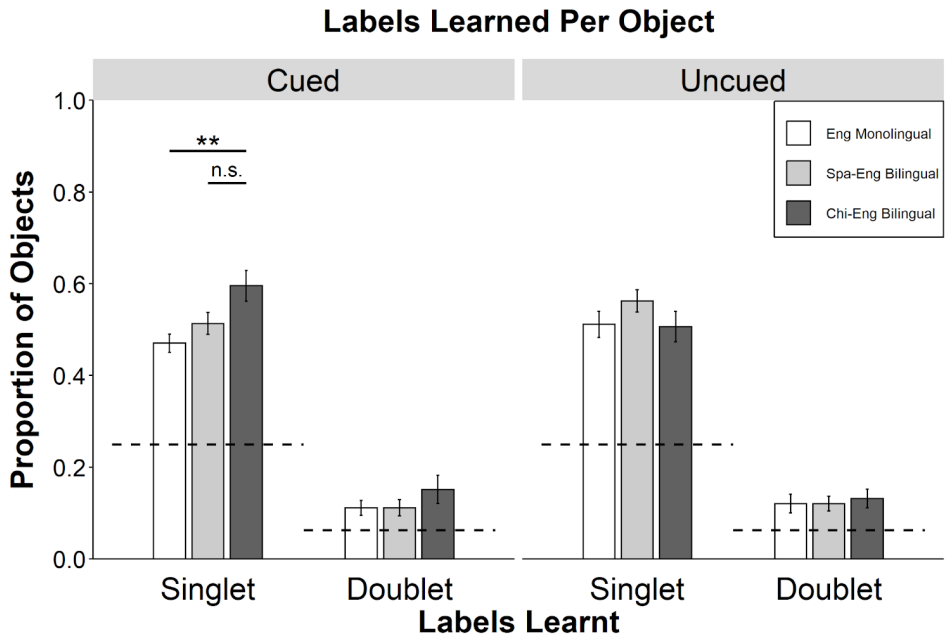
According to the pre-registered analysis plan, we also examined how participants learned the words for referents in each condition. Specifically, successful learning could be achieved by predominantly learning one label for each referent, predominantly learning two labels for each referent, or a mix of both (Benitez et al., 2016; Benitez & Li, 2022; Ichinco et al., 2009). By using one-sample t-tests, we compared the proportion of learning singlets and doublets against chance (chance for singlet = ¼ = .25; chance for doublet = ¼ × ¼ = .0625). Results showed that average learners learned singlets above chance ($M = .52$, $SD = .19$, $t(148) = 23.23$, $p < .001$, $d = 1.90$) and doublets above chance ($M = .14$, $SD = .14$, $t(148) = 6.97$, $p < .001$, $d = .57$).

1.4 Three-Way ANOVA for Learned Objects as a Function of Labels (Singlet vs Doublet), Condition, and Group

According to the pre-registered plan, we conducted a three-way mixed ANOVA to test the effect of label (singlet and doublet), condition, and group on the proportion of learned objects. **Figure S4** below displays the proportion of objects (out of 8) for which learners acquired a single label (singlet) or both labels (doublet) by condition and group.

Figure S4

Learned Objects (in Proportion) as a Function of Label, Condition and Language Group.



Note. Mean proportion of objects (and standard error of the mean) for which one label or two labels were learned in each condition and language group. Asterisks denote significant between-group differences (*$p < .05$, **p < .01, n.s. non-significant). Dashed line denotes chance performance (0.25 for singlet and 0.0625 for doublet).

Results showed a significant main effect of Label, such that there was a higher proportion of objects for which learners acquired singlets ($M = .52$, $SD = .19$) than doublets ($M = .12$, $SD = .14$; $F(1, 146) = 624.98$, $p < .001$, $\eta^2 = .59$). The main effect of Group was significant ($F(2, 146)$

= 3.34, p = .038, η^2 = .01) and the interaction between Group and Condition was also significant (F(2, 146) = 5.83, p = .004, η^2 = .01).

Post-hoc analysis on the Group×Condition interaction on each level of label showed that language groups differed only in the Cued condition in learning singlets (F(2, 146) = 5.73, p_{adj} = .016, η^2 = .07), with p-value Bonferroni adjustments. Specifically, Chinese-English bilinguals were significantly more likely to learn singlets (M = .60, SD = .21) than English monolinguals (M = .47, SD = .15, p_{adj} = .003) and numerically more than Spanish-English bilinguals (M = .51, SD = .18), though this effect did not reach significance (p_{adj} = .080). The groups, however, did not differ in learning doublets in the Cued condition [F(2, 146) = 1.07, p_{adj} = 1.000, η^2 = .01, English monolinguals (M = .11, SD = .12), Spanish-English bilinguals (M = .11, SD = .13), Chinese-English bilinguals (M = .15, SD = .19)]. No group differences were found in the Uncued condition when learning singlets [F(2, 146) = 1.25, p_{adj} = 1.000, η^2 = .02, English monolinguals (M = .51, SD = .21), Spanish-English bilinguals (M = .56, SD = .18), Chinese-English bilinguals (M = .51, SD = .21)] nor learning doublets [F(2, 146) = .10, p = 1.000, η^2 = .00, English monolinguals (M = .12, SD = .15), Spanish-English bilinguals (M = .12, SD = .12), Chinese-English bilinguals (M = .13, SD = .12)]. The results converge to show that Chinese-English bilinguals' advantage in word learning did not necessarily lie in learning both labels to an object. Instead, such an advantage was manifested in learning either label of an object (but not both) when a lexical tone cue was present.

2 Other Analyses

2.1 Considering Age and Education for the Effects of Condition and Group

As we observed a significant interaction between Condition and Group in the main text, we additionally considered possible confounds of subjective characteristics (Age and Education

level), as Chinese-English bilinguals were older than the other two groups, and had more participants with higher education (see the report in the Method section).

First, we ran two additional GLMM models to include Age (a continuous factor) and Education (a categorical factor) into the prior main model, respectively. By adding Age, results showed that the Condition×Group interaction still held (Wald X^2 (2) = 7.14, p = .028), with a significant main effect of Age (Wald X^2 (1) = 3.94, p = .047). By adding Education, the Condition×Group interaction still held (Wald X^2 (2) = 7.14, p = .028), with a non-significant main effect of Education (Wald X^2 (6) = 7.93, p = .243).

Second, we also ran the same GLMM analyses in the main text with fixed effects of Condition and Group (1) after excluding the participants with advanced education (n = 11, Masters or Ph.D. or higher; all Chinese-English bilinguals) and (2) after excluding the participants who were comparatively older (older than or equal to 28 years old) (n = 13, 3 Spanish-English bilinguals, and 10 Chinese-English bilinguals). Results from the subsets after deleting participants with advanced education (Condition×Group interaction, p = .019) and after deleting older participants (Condition×Group interaction, p = .037) were consistent with those from the full dataset.

These additional analyses on age and education level converge to provide strong evidence of a robust Condition×Group interaction reported in the main text, even controlling for the subjective characteristics.

SECTION 6 CONFIDENCE IN LEARNING

The section below does further analysis on the relation between one's confidence in learning and their actual SWL performance. Data analysis and the results output are presented below.

We paired each participant's confidence in learning with their actual SWL performance, in the Cued and the Uncued condition respectively, and conducted a series of simple linear regression models. We had predictors of the subjectively rated confidence in learning (0-5, 0 being not learning at all, and 5 being learning a lot), language groups, and their interaction; we had the criterion of SWL performance in the Cued and Uncued conditions respectively. Results showed that confidence significantly predicted SWL performance in the Cued condition, but not that in the Uncued condition. **Table S8** below denotes the model outcome, and **Figure S5** below depicts the individual data points and regression lines by language groups.

As for Model 1 where the SWL performance in the Cued condition was the criterion, the regression model was significant (R^2 = .13, adjusted R^2 = .10, $F(5,143) = 4.28$, $p = .001$). The coefficient confidence significantly contributed to Model 1 and predicted SWL performance (p = .009). The between-group differences were consistent with the results in the main text that Chinese-English bilinguals outperformed English monolinguals (p = .004) and Spanish-English bilinguals (p = .001) in the Cued condition. Besides, the Group×Confidence interaction was not significant (p > .508), suggesting that the degree of predicting SWL performance from confidence did not differ by language group in the Cued condition.

As for Model 2 where the SWL performance in the Uncued condition was the criterion, the model was not significant (R^2 =.02, adjusted R^2 = -.01, $F(5,143) = .56$, p = .732). In contrast, the coefficient confidence did not contribute to the model and did not predict SWL performance (p = .242). Language groups did not differ in predicting SWL performance (p > .462), consistent

with the findings in the main text that language groups did not differ in learning in the Uncued condition. Furthermore, the Group×Confidence interactions were not also significant ($p > .593$), suggesting that the slope of predicting SWL performance from confidence did not differ by language group in the Uncued condition. In short, learners' confidence in learning, reported retrospectively after completing the word learning tasks, significantly predicted actual statistical word learning performance when a linguistic cue was presented but not when without such a cue. Language groups did not differ in the propensity of linking one's confidence in learning and the SWL performance.

Table S8

Simple Linear Regression Models for Confidence in Learning and SWL

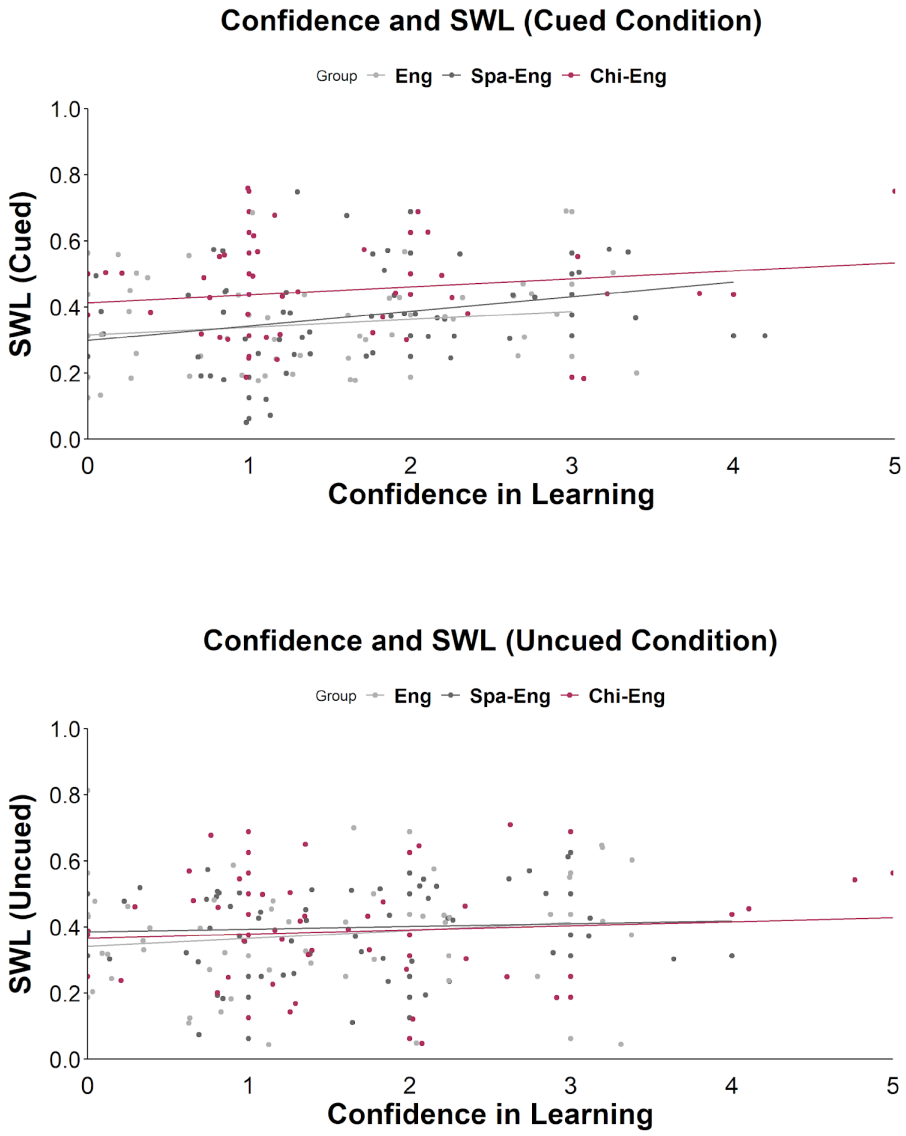
Model 1: SWL in the Cued Condition and Confidence				
Coefficients	*b*	*SEb*	*t*	*p*
Confidence	.03	.01	2.65	**.009**
Spa – Eng	.01	.03	.48	.637
Chi – Eng	.10	.03	3.32	**.001**
Chi – Spa	.08	.03	2.91	**.004**
(Spa – Eng)*Confidence	.02	.03	.76	.449
(Chi – Eng)*Confidence	.00	.03	.02	.986
(Chi – Spa)*Confidence	.02	.03	.66	.508
Model summary	R^2	ΔR^2	*F*	*p*
	.13	.10	4.28	**.001**
Model 2: SWL in the Uncued Condition and Confidence				
Coefficients	*b*	*SEb*	*t*	*p*
Confidence	.01	.01	1.18	.242
Spa – Eng	.02	.03	.74	.462
Chi – Eng	.01	.03	.25	.801
Chi – Spa	-.01	.03	-.41	.679
(Spa – Eng)*Confidence	-.02	.03	-.54	.593
(Chi – Eng)*Confidence	-.01	.03	-.37	.710
(Chi – Spa)*Confidence	.00	.03	-.13	.896
Model summary	R^2	ΔR^2	*F*	*p*
	.02	-.01	.56	.732

Note. The bold numbers stand for the significant *p*-values for the significant predictor of Confidence, and/or the significant regression models, two-tailed. We used dummy coding for groups and changed the reference groups to

get the above values. The table suggests that confidence of learning did predict SWL in the Cued condition, but not in the Uncued condition.

Figure S5

Simple Linear Regression Models for Confidence in Learning and SWL



Note. The figure depicts how confidence in learning (continuous predictor from 0 to 5) and language groups predicted SWL performance in the Cued condition (**Top Panel**) and that in the Uncued condition (**Bottom Panel**) in two interactive simple linear regression models. Straight lines denote the slope of predictability–how confidence

predicts SWL performance–for English monolinguals (in light grey), for Spanish-English bilinguals (in dark grey), and for Chinese-English bilinguals (in maroon). Colorful dots denote individual data points for the three groups.