Cognition 213 (2021) 104788

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Asymmetric memory for birth language perception versus production in young international adoptees☆

Wencui Zhou[a], Mirjam Broersma[b,*], Anne Cutler[c]

[a] Tilburg University Language Center, Professor Cobbenhagenlaan 205, 5037 DB Tilburg, the Netherlands
[b] Centre for Language Studies, Radboud University, 6500 HD Nijmegen, the Netherlands
[c] MARCS Institute and ARC Centre of Excellence for the Dynamics of Language, Western Sydney University, NSW 2751, Australia

ARTICLE INFO

Keywords:
Adoption
Language retention
Speech perception
Speech production
Segments
Lexical tones

ABSTRACT

Adults who as children were adopted into a different linguistic community retain knowledge of their birth language. The possession (without awareness) of such knowledge is known to facilitate the (re)learning of birth-language speech patterns; this perceptual learning predicts such adults' production success as well, indicating that the retained linguistic knowledge is abstract in nature. Adoptees' acquisition of their adopted language is fast and complete; birth-language mastery disappears rapidly, although this latter process has been little studied. Here, 46 international adoptees from China aged four to 10 years, with Dutch as their new language, plus 47 matched non-adopted Dutch-native controls and 40 matched non-adopted Chinese controls, undertook across a two-week period 10 blocks of training in perceptually identifying Chinese speech contrasts (one segmental, one tonal) which were unlike any Dutch contrasts. Chinese controls easily accomplished all these tasks. The same participants also provided speech production data in an imitation task. In perception, adoptees and Dutch controls scored equivalently poorly at the outset of training; with training, the adoptees significantly improved while the Dutch controls did not. In production, adoptees' imitations both before and after training could be better identified, and received higher goodness ratings, than those of Dutch controls. The perception results confirm that birth-language knowledge is stored and can facilitate re-learning in post-adoption childhood; the production results suggest that although processing of phonological category detail appears to depend on access to the stored knowledge, general articulatory dimensions can at this age also still be remembered, and may facilitate spoken imitation.

1. Introduction

Learning the native language is arguably the child's most important task, and even in the first year of life enormous progress towards this goal is achieved. Before age 1, infants have learned to distinguish most of the phonetic contrasts used in the native language, have begun compiling an initial vocabulary whereby they can recognise dozens of spoken words (Johnson, 2016) and acquiring their native language grammar (Gervain, Nespor, Mazuka, Horie, & Mehler, 2008), and are ready to start trying out their own spoken communicative skills (Johnson, 2016).

For some individuals, however, this initial work needs to be started all over again and repeated with another language. This is the lot of international adoptees, whose language in the first months of life is usually not the language of the country which becomes their adoptive home. After adoption, they start again on the task of learning a new set of contrasts, a new vocabulary, and new speech production targets.

This repeat performance is a very achievable task, in fact, as has been comprehensively documented in a significant series of studies by Roberts and colleagues (Roberts, Krakow, & Pollock, 2003; Roberts et al., 2005; Krakow & Roberts, 2003, Krakow, Tao, & Roberts, 2005; Scott, Roberts, & Krakow, 2008; Scott, Pollock, Roberts, & Krakow, 2013) of a similar target population to that involved here, namely young adoptees from China who learned English in their new homes. Although there is typical individual variation, these studies found that adoptees acquire the new environmental language with speed and efficiency, such that within two years 90–95% of them are performing at or above average on standardized speech-language tests normed on monolingual English speakers. By school age their phonological processing, spoken language

☆ This paper is a part of special issue "Special Issue in Honour of Jacques Mehler, Cognition's founding editor".

* Corresponding author.
E-mail addresses: w.zhou_2@uvt.nl (W. Zhou), m.broersma@let.ru.nl (M. Broersma), a.cutler@westernsydney.edu.au (A. Cutler).

https://doi.org/10.1016/j.cognition.2021.104788
Received 28 October 2020; Received in revised form 18 May 2021; Accepted 20 May 2021
Available online 2 July 2021

comprehension and reading skills are likewise at or above average, and in line with those of their classmates.

In contrast, the language acquired pre-adoption (hereafter, the "birth language") is apparently forgotten (Isurin, 2000; Nicoladis & Grabois, 2002), and by adulthood, as many studies have shown, adoptees retain no conscious memory of the birth language and are to all effects and purposes native speakers of the language of their adoptive land. This can be such a strong identification that they may resist suggestions to the contrary (Ventureyra, Pallier, & Yoo, 2004).

Notwithstanding this, however, teenage and adult adoptees have, without awareness, retained knowledge of the birth language – at least, of its phonology. This does not confer any performance benefits when the task is one of simple discrimination (Choi, Broersma, & Cutler, 2015; Ventureyra et al., 2004; Werker, 1986), although even in that case the birth language can be seen to activate language processing areas in the brain of adoptees (but not of control participants without the relevant early experience; Pierce, Klein, Chen, Delcenserie, & Genesee, 2014; Pierce, Chen, Delcenserie, Genesee, & Klein, 2015). Performance on an identification task (more difficult than simple discrimination) indeed exhibits a relearning advantage for those with the birth language knowledge however, independently of the number of months or years of a participant's birth language exposure (Bowers, Mattys, & Gage, 2009; Singh, Liederman, Mierzejewski, & Barnes, 2011; Choi, Broersma, & Cutler, 2017, Choi, Cutler, & Broersma, 2017). This suggests that the stored knowledge is abstract in nature rather than being an accrual of exposure episodes, and this conclusion is supported by the fact that perceptual relearning benefits both generalise within perception from trained to untrained structures, and also transfer from perception to production (Choi, Cutler, & Broersma, 2017).

Note that this stored early knowledge does not appear to benefit language users beyond the particular case of relearning the ability to identify birth-language sounds. Children growing up with more than one language have been reported to show evidence of some cognitive advantages from infancy onward (Kovács, 2009; Kovács & Mehler, 2009), and later display advantages in performing word learning tasks (Kaushanskaya & Marian, 2009), recognising prosodic structure (Stepanov, Pavlič, Stateva, & Reboul, 2018), detecting talker change (Levi, 2018) and learning novel phonological contrasts (Tremblay & Sabourin, 2012). But there is no equivalent phonological processing gain from the sequential exposure to two languages, as experienced by adoptees, since adoptees show no performance advantage at all with a novel contrast, only with the re-encountered birth language contrasts (Choi, Broersma, & Cutler, 2018). Bilingual advantages are held to stem from practice in executive control as part of the effort of choosing between two languages (two sets of names, two options for word order and so on; see Antoniou, 2019, for an overview of this as yet unresolved debate). Adoptees are not bilinguals but, rather, sequential monolinguals, and their situation does not exercise such choice processes.

Nonetheless, abstract knowledge about the birth language has been stored, and under the right circumstances can be tapped even decades later in adulthood. Choi, Broersma, and Cutler (2017) showed that this knowledge would have been in place to at least some degree even for adoptees whose new language exposure commenced in the first half year of life. As even decades of separation from the birth language do not result in the loss of such stored, abstract knowledge, this raises the question of how available it is to young children who have been adopted much more recently. On the one hand, one might expect that stored birth language knowledge is readily available to them, as it is relatively recent. On the other hand, the accessibility of birth language knowledge might decrease rapidly, such that even young children require re-exposure to be able to access it. Finally, one might ask if children can indeed draw at all on abstract knowledge of the birth language, in the way that adults have been shown to do, or whether such access may depend upon greater linguistic maturity. The evidence cited above on adoptees' ready acquisition of the new language, and apparent loss of conscious access to the birth language, cannot answer these questions because it does not include measures that tap into knowledge held without awareness. Most studies showing the presence of such birth-language knowledge have concerned adults of at least student age. Only two studies concerned children, and in both cases the tested populations were in their second decade of life. First, the cited Pierce et al. (2014, 2015) study tested 21 adopted children from China at an average age of 13 and a half years, and found that the brain activation of the adoptees, unlike those of controls who had not been previously exposed to Chinese, resembled those of Chinese non-adopted children. Second, the Singh et al. (2011) study tested eight adopted participants from India at an average age of 12 years four months, and found that after training, adoptees discriminated a contrast from their birth language better than a group of non-adopted controls whose native language was English.

In the present study, we address this question by testing children who had been adopted for more than one but less than 10 years, from an age between 9 months and four and a half years, at an average age at testing of seven and a half. Note that we will not be able to tease apart effects of adoption age, time since adoption, and age at testing, as they are inherently interrelated, with the former two adding up to the latter. We test 46 adopted children and compare their performance on several language tasks to that of 47 age-matched control children who share their adopted native language, and another 40 who share their birth language. The language tasks assess their memory for words of their birth language (as in the studies of birth language loss, e.g. Isurin, 2000), and their ability to learn to perceive and to imitate a segmental phonetic distinction that does not feature in the adopted language (the type of learning advantage that has been demonstrated in most of the studies of older children and adults with an adoption background), as well as their ability to perceive and to imitate a phonological contrast in a dimension (lexical tone) that the adoptive language does not use (as in the work of Pierce et al., 2014, 2015). As well as testing these *perceptual* learning abilities, we assess the native-speaker identification and rating of their spoken *productions* of the same segmental and tonal contrasts.

Our predicted answer to the question of these children's performance in comparison to that of older children and adults is that, based on the evidence so far, we will find that the adopted children will have entered the state that has been observed for adults as a function of having become users of the environmental language only. There should then be no observable gradient of accessibility of the stored phonological knowledge. We predict that their conscious access to their birth language will have been lost, resulting in an inability to perform better than the adopted-language controls on vocabulary knowledge. We also predict no difference from the same null-exposure control group in perceptual performance *before* training exposure; but we predict a significant adoptee-control difference *appearing with training*. In other words, these child adoptees will respond to training in the same way as was observed with adult adoptees; from an equal starting position to the adopted-language controls, they will be enabled by the stored abstract knowledge to outstrip the controls in learning a perceptual distinction. Further, we predict that such abstract knowledge will buttress the transference of successful perceptual distinctions to production of the same contrasts as well, resulting in the identification and rating of adoptees' productions again surpassing those of the null-exposure controls. As relearning advantages for adoptees have been shown for segmental contrasts but have not been investigated for tones (Pierce et al., 2014, 2015, did not investigate relearning), we have no reason to expect a difference between our segmental and tonal contrasts; we extend previous research by investigating both.

Finally: our adoptees were all born in China and became residents of The Netherlands. China has many languages. The two largest language communities are Mandarin and Cantonese, and these two languages were also the most common birth languages of adoptees who came from China to The Netherlands. Our participant population contained near-equal numbers from each language. Thus we selected separate target sounds and tones for each language, constructed separate sets of test materials for each language, and had separate control groups and native test populations for each language. This means that we effectively report two separate experiments, one in each language. Since in every procedural respect each group was treated identically, however, we avoid repetition by reporting the Cantonese and Mandarin studies together.

2. Experiments

2.1. Participants

All adoptee and non-adopted Dutch control participants were young schoolchildren in the Netherlands who received a small reward for their voluntary participation. All were reported by their parents to have normal speech, hearing, and motor control, and normal or corrected-to-normal vision (seven of the 46 adoptees had an impairment that did not affect their task performance; see Zhou, 2015, for detail).

The adoptees had all been born in China and raised there for at least their first 9 months, had been adopted by Dutch-speaking families in the Netherlands as infants or young children, and at the time of testing had lived in the Netherlands for on average 5 years 3 months, minimally 13 months. For each adoptee group there was a non-adopted Dutch control group (born and raised in the Netherlands in their Dutch-speaking birth families) and a non-adopted Chinese control group (born and raised in China in birth families with the relevant language), with each three groups of children being matched in age, in gender, and in music training outside school. No children in either of the adoptee or Dutch control groups had previously received any Chinese language training (after adoption, in the adoptee case), and none were reported to be able to understand any Chinese. All spoke Dutch fluently, as was confirmed using the Dutch MacArthur-Bates Communicative Development Inventory (Zink & Lejaegere, 2002).[1] The adoptee and Dutch control groups for each language were also matched on parents' educational level (complete data on all matching factors can also be found in Zhou, 2015). Informed written consent was obtained from participating children's parents, and from participating adults.

The *Cantonese group* consisted of 21 Cantonese adopted children (15 female; age: 4y 5mo – 10y 10mo, mean 7y 7mo, SD 1y 10mo; adoption age: 9mo – 4y 6mo, mean 2y 1mo, SD 1y 1mo; time since adoption: 1y 10mo – 9y 11mo, mean 5y 5mo, SD 2y 4mo). Their non-adopted Dutch control group contained 22 non-adopted children (12 female; age: 4y 7mo – 10y 7mo, mean 7y 6mo, SD 1y 9mo), recruited from the siblings of the Cantonese adoptees (5), and through informal networks (17). Their non-adopted Chinese control group contained 22 fluent Cantonese-Mandarin bilingual children (13 female; age: 4y 3mo – 10y 5mo, mean 7y 3mo, SD 1y 11mo) in Guangzhou City, whose home language was Cantonese. Note that Cantonese children typically become bilingual when they enter school, and that in school populations there are thus very few Cantonese monolingual children in China. Whereas the development of phonological representations might follow different trajectories in bilingual vs monolingual children, such differences would be expected to have been resolved before the age at which our participants were tested (Burns, Yoshida, Hill, & Werker, 2007; Ramon-Casas, Swingley, Sebastián-Gallés, & Bosch, 2009).

The *Mandarin group* contained 25 Mandarin adoptees (15 female; age: 4y 1mo – 10y 10mo, mean 7y 4mo, SD 1y 9mo; adoption age: 10mo – 4y 2mo, mean 2y 2mo, SD 11mo; time since adoption: 1y 1mo – 9y 10mo, mean 5y 2mo, SD 2y). Their 25 non-adopted Dutch controls (12 female; age: 4y 2mo – 10y 4mo, mean 7y 1mo, SD 1y 9mo) were recruited from the siblings (5) and relatives (3) of the Mandarin adoptees, and through informal networks (14). Their 18 non-adopted Chinese controls (7 female; age: 4y 1mo – 8y 3mo, mean 6y 6mo, SD 1y 4mo) were Mandarin monolinguals in Beijing.

Further, four native-speaker groups were recruited to assess the participants' production: two of Cantonese-native university students in Guangzhou (Identification test: N = 24, Rating test: N = 21; age 19–25 years, M = 22 years, 23 females), and two of Mandarin-native university students in Beijing (Identification test: N = 22, Rating test: N = 22; age 19–27 years, M = 23 years, 22 females).

2.2. Materials

For each language we selected a segmental contrast of affricate consonants, and a tone contrast. Dutch contrasts include neither these segments (Gussenhoven, 1992), nor any tones, so all were expected to be difficult for Dutch participants. Tones, which native learners of tone languages generally acquire later and with more uneven learning patterns than segments (Götz, Yeung, Krasotkina, Schwarzer, & Höhle, 2018), were expected to prove somewhat more difficult.

The *Cantonese affricate contrast* involved the alveolar affricates [t͡s] and [t͡sʰ], which contrast in aspiration (Bauer & Benedict, 1997). These are fairly difficult sounds which are reported to be produced relatively late by Cantonese children (So & Dodd, 1995); nonetheless, infants are likely to have been exposed to the sounds as they occur in early vocabulary (e.g., the words for *bed* and *banana* contain [t͡s], while *mouth* and *outside* contain [t͡sʰ]). (Note that Dutch, unlike English, does not use aspiration to distinguish stop consonants.)

The *Mandarin affricate contrast* involved the retroflex affricates [tʂ] and [tʂʰ] which again are distinguished in aspiration (Duanmu, 2007). These are reported to emerge in infant Mandarin productions comparatively early (Hua & Dodd, 2000); again they appear in early vocabulary items (*porridge, eat, bed* etc.).

The *Cantonese tone contrast* involved Tones 2 and 5, which are both rising tones, distinguished by rising offset height, with the offset of Tone 2 being higher than that of Tone 5. Both are reported to be mastered during the second year, with Tone 2 appearing earlier than Tone 5 (So & Dodd, 1995). This ordering may reflect frequency (Tone 2 is more frequent), or perceptual conflation (they are the only two rising tones of the six Cantonese tones), or both.

The *Mandarin tone contrast* was between Tones 2 and 3; here Tone 2 is a smooth rise, while Tone 3 consists of two movements, falling and rising. Both tones are typically acquired before the age of three (Hua & Dodd, 2000), and here too there is asymmetry in order of acquisition (Tone 2 slightly precedes Tone 3 in production; Zhu, 2002), and frequent confusion of the two in early stages of native acquisition (Tsao, 2008). Again, the fact that both these two Mandarin tones have a rise at their end appears to make them perceptually similar.

For each language, there were 64 items, forming 32 minimal pairs of disyllabic pseudowords, 16 per contrast (see supplementary materials). (As all phonotactically legal monosyllables in Chinese languages are existing words, the disyllabic pseudowords effectively consisted of real words in the same way as an English pseudoword *doky* would consist of *dough* and *key*). For all stimuli, the first syllable was always [a] with tone 1 (High-Level in each language), followed by a second syllable realising a contrast, which was at the onset in affricate stimuli, and on the vowel for tone stimuli. The structure of the second syllable (with C for consonant, V for vowel, and G for glide) was C-V, C-V-C, C-G-V, or C-G-V-C. Onset consonants were plosives, fricatives, nasals, liquids, or glides; coda consonants were nasals. Apart from in the segmental contrasts, no affricates were used, and apart from in the tonal contrasts, all syllables bore tone 1.

For each language, three female native speakers (aged 21–29 years) each recorded the 64 stimuli (16 minimal pairs per contrast) in a sound-

[1] Parents filled in the highest level of the Dutch vocabulary checklist, N-CDI 3, designed for children aged 30–37 months, which consists of four parts: 1) active vocabulary, 2) phrases, 3) length of sentences produced, and 4) questions on comprehension, semantics, and syntax. Raw scores of each part were converted to percentiles based on normative tables (Zink & Lejaegere, 2002), and averaged per child. All children had an average score above the median (Cantonese adoptees: mean 89, SD 13; non-adopted Dutch controls: mean 93, SD 5; Mandarin adoptees: 87, SD 21; non-adopted Dutch controls: 94, SD 7), most of them near ceiling (in line with Snedeker, Geren, & Shafto, 2007). *t*-tests showed no statistical differences between adoptees and controls ($ps > 0.05$).

proof booth in a clear citation style (sample rate: 44.1 kHz, 16-bit). The number of tokens that they recorded varied from 1 to 11 across stimuli, depending on which tasks they were used for (see supplementary materials). For each contrast, 12 of the minimal pairs (24 items) were used for the perception training, with 10 tokens per item per speaker (namely 1 token per item per speaker for each of the 10 training blocks). Further, for each contrast, all 16 minimal pairs (32 items) were used for the perception tests, with one token per item per speaker. Finally, for the production test, 2 of the minimal pairs (4 items) were used for each contrast, produced by two of the speakers, with one minimal pair per contrast per speaker, and two tokens per item. Training, perception tests, and production tests all used unique tokens that were not used anywhere else in the study; the same tokens were used at pretest and posttest.

2.3. Procedure

Each non-adopted Dutch and Chinese control group received the same training and test as its relevant adoptee group. All tasks involving adoptees and controls were carried out during four visits to each participant's home within a two-week period. There were 10 perceptual training blocks, and their effect was tested in perception and production tests at the beginning of visit 1, before any training (pretest), and at the end of visit 4, after the 10 training blocks (posttest). Each training block, perception test, and production test contained one part for affricates and one for tones (in that order[2]). Each training block contained 24 trials per contrast, each perception test 16, and each production recording four trials per contrast. Each test and the first training block per visit began with instructions (in Dutch for participants in the Netherlands, and in Cantonese and Mandarin, respectively, for the non-adopted Chinese control groups) including opportunity to practice, delivered as part of the game, and three visits also included a test of Chinese vocabulary.

2.3.1. Vocabulary test

A picture-matching task was constructed to test for residual vocabulary knowledge. Test items were 36 simple picturable referents (e.g., *banana*, *smile*, *shoes*); each was paired with a distractor picture, for 72 pictures in all. The pictures/referents were the same for each language group, with the Mandarin test words all in the Mandarin CDI and reported to be understood by >50% of 12-months-olds (Hao, Shu, Xing, & Li, 2008). At each of three testing occasions (first, second and fourth visit), 12 items were presented. The task was to pick the named referent from two pictures; the presentation used the cartoon framework described in the following section.

2.3.2. Perception testing and training

All participants were trained on the perception of an affricate contrast and a tone contrast. Stimuli were different for the Cantonese and Mandarin language group. In each language, the perception tests and the training used an XAB non-word discrimination task. Participants first heard a non-word stimulus (X), and then two other non-word stimuli (A and B), all produced by different speakers. The first stimulus (X) was presented at 1000 ms after onset of the trial. There was an inter-stimulus interval (ISI) of 1500 ms between the offset of the first and the onset of the second stimulus (i.e., X and A), and an ISI of 1000 ms between the offset of the second and the onset of the third stimulus (i.e., A and B). Previous work suggests that ISIs of this length motivate phonological processing, in contrast to shorter ISIs which stimulate phonetic processing (Werker & Logan, 1985). The participants' task was to judge which of the last two stimuli (A or B) was identical to the first stimulus (X). Each of the two affricates and each of the two tones occurred as X in the XAB task equally often, and A and B were correct (matching X) equally often, in each block.

The XAB task was presented as a cartoon game, with a mother animal who spoke the stimulus X and two baby animals (standing to each side of the mother animal) who spoke the stimuli A (first, participants' left side) and B (second, participants' right side). Each animal was consistently combined with the same speaker. Participants were instructed to indicate which baby animal correctly imitated the mother animal. They did this by pressing a button on a button box, the left button (stimulus A) or the right button (stimulus B), after the second animal had spoken. During training, participants received visual feedback about the correctness of each response. An example screen is depicted in Fig. 1. During the test, there was no feedback on the correctness of the response, but motivation videos were played as an incentive after each response (see below).

All videos showed three animals – dinosaurs for affricate contrasts, pandas for tone contrasts[3] – against a variety of backgrounds, decorations and objects. In training, feedback on the correctness of the response was given: After a correct response the chosen baby animal jumped up and down with colorful stars above its head; after an incorrect response the baby animal cried and rubbed its face. For the test, the screen showed the test video (2/3 of the screen horizontally) with all three animals, plus the motivation video (the other 1/3 of the screen), showing a baby animal jumping up one step on a staircase after each button press (regardless of response correctness). After every eight trials, the baby animal reached the top of the staircase, and opened a gift box from which a unique gift popped out, accompanied by stars, a balloon and cheerful sound effects.

The perception tests yielded a total of 5952 observations (Cantonese: 43 participants * 2 test points [pretest and posttest] * 2 contrasts * 16 trials = 2752; Mandarin: 50 participants * 2 test points * 2 contrasts * 16 trials = 3200).

The experiment was conducted using Presentation software (Neurobehavioral Systems, version 14.7) on a laptop computer (HP EliteBook 8540P with resolution 1366 × 768 pixels). Participants were seated at a comfortable viewing distance from the laptop screen. They heard the auditory stimuli through Sennheiser HD 280 headphones and responded using a button box with two buttons (MPI Dual Button box: Serial port via USB, Baudrate 38400, Data-8bit, and StopBit-1). The production responses were recorded via a Roland EDIROL R-09 recorder.

2.3.3. Production testing and assessment

For the production recordings, an imitation task was used involving two speakers (representing the two baby animals in the videos). On each trial, participants heard a target word repeated twice (two tokens from the same speaker; ISI 1500 ms), and a Dutch instruction to repeat the word (*'nu jij', it's your turn*). The test videos resembled the training, except that the baby animal uttered the target words into a microphone, and the mother animal pointed another microphone towards the participant.

For the production assessment, native Cantonese and Mandarin listeners in Guangzhou and Beijing respectively identified and rated the adoptees' and non-adopted Dutch controls' recordings. In the *identification* test (two-alternative forced choice), a computer screen showed a minimal pair in Chinese characters, differing in the crucial affricate or tone (blocked by contrast), as one recording was played (ISI 1 s.). Participants responded by clicking on the word they heard. In the *rating* test the screen showed a single target word, and one recording was played (ISI 1 s). Participants rated how similar it sounded to the target word on a 4-point scale (1 = quite different "完全不一样", 2 = a little similar "有

[2] The order of the contrasts was fixed rather than counterbalanced to prevent an unbalanced data set in the event of participants dropping out, although as it turned out all participants in fact completed the study.

[3] As with the order of presentation of the contrasts, the combination of contrasts and animals was fixed rather than counterbalanced in case of participant drop-out.

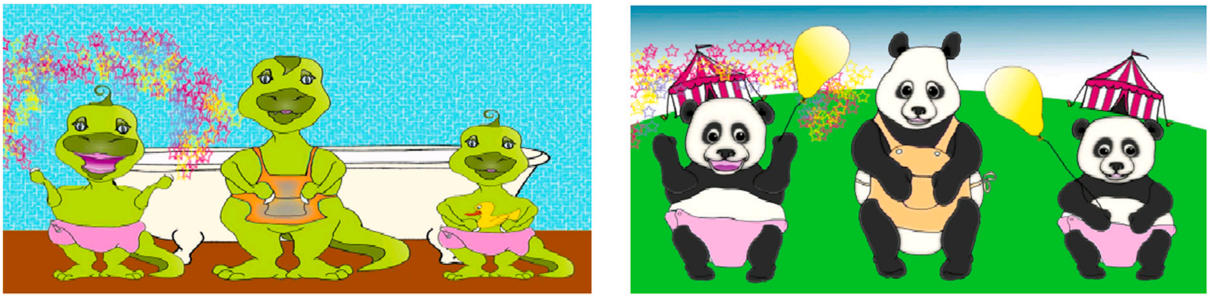

Fig. 1. Example of a positive feedback video with the dinosaur family (left), and with the panda family (right).

点相似", 3 = very similar "非常相似", 4 = identical "完全一样". A scale with an even number of steps was used to avoid excessive use of the middle step). Recordings of the non-adopted Chinese controls were not included, as we expected their inclusion to diminish the use of the higher end of the scale for the adoptees' and non-adopted Dutch controls' recordings.

For this assessment, listeners were tested in small groups in a quiet room, in two sessions over two days. Each heard (through headphones) all the recordings of the adoptees and non-adopted Dutch controls (Cantonese: 43 participants * 2 test points * 2 contrasts * 4 items = 688; Mandarin: 50 participants * 2 test points * 2 contrasts * 4 items = 800). The presentations were blocked by contrast (one per session) with order counterbalanced. Listeners performed either the identification or the rating task.

2.4. Results

2.4.1. Vocabulary

All non-adopted Chinese control participants scored 100% in this test, confirming the task's validity, and their results were not further analysed. The performance of adoptees and non-adopted Dutch controls on the other hand was close to 50% correct, which constitutes chance level (Cantonese: adoptees 54%, non-adopted Dutch controls 51%; Mandarin: adoptees 52%, non-adopted Dutch controls 49%).

For the adoptees and non-adopted Dutch controls, responses with reaction times (RT) more than two SDs from a participant's mean RT were removed (Cantonese: 35/1548 = 2.3% of the data; Mandarin: 37/1800 = 2.0%; similar for adoptees and Dutch controls), and separate analyses of variance (ANOVAs) by participants (F_1) and by items (F_2) were conducted for each language with the proportion of correct responses as dependent variable and Group (adoptees, non-adopted Dutch controls) as independent variable. (We have collapsed across the three test sessions because different items were used in each session.) There was no significant effect of Group (all $ps > 0.05$). Further, one-sample *t*-tests showed that the participants failed to score differently from chance (all $ps > 0.05$). These results thus show that the adoptees and Dutch controls had no knowledge of Chinese vocabulary.

2.4.2. Perception

Perception tests consisted of two blocks of 8 trials (one for affricates, one for tones). If a participant pressed the same key throughout a block (except in cases involving more than three successive correct responses), we took this as a sign of failure either to understand the task or to concentrate, and excluded those 8 trials from analysis (for a similar procedure, see Goriot, McQueen, Unsworth, Van Hout, & Broersma, 2020). Responses with RTs more than two *SDs* from a participant's mean RT were also removed. In total, this removed 11% of the total data (653/5952 trials; almost equally from the Cantonese and Mandarin sets); thus after the data cleaning 89% of trials (5297) were entered into the analyses reported below.

To analyse the perception tests, ANOVAs were done in two ways, treating participants (F_1) and items (F_2) as random variables, to take into account the variance across participants and across stimuli, respectively. Effects were only considered significant if the associated *p*-value was <0.05 in both analyses. Mauchly's Test of Sphericity was used to test for violations of the assumption of sphericity; if needed, degrees of freedom were adjusted using Greenhouse-Geisser estimates of Sphericity.

The Chinese non-adopted control participants scored, as predicted, very high on all contrasts at all test points (Cantonese: affricates, pretest: 93% correct, posttest: 96%; tones, pretest: 75%, posttest: 87%; Mandarin: affricates, pretest: 90%, posttest: 94%; tones pretest: 93%, posttest: 90%). Scores on the Cantonese tone decisions were a little lower, potentially related to the Cantonese children's bilingualism, but certainly consistent with the perceptual conflation that is reported to occur between these two rising tones among the six Cantonese tones. The non-adopted Chinese control participants outperformed the two Netherlands groups across the board (see Fig. 2). ANOVAs by participants and by items, with proportion of correct responses as dependent variable and Group and Test point as independent variables, confirmed that for each language and for each contrast, the non-adopted Chinese controls significantly outperformed both the adoptees and the non-adopted Dutch controls (all $ps < 0.001$). The Chinese control data were collected in order to provide assurance that the task that our participants undertook was within the capabilities of the tested age group. This indeed proved to be the case, and the results of the control participants will not be discussed further.

For adoptees and their non-adopted Dutch controls, mean percentages correct on the pre- and posttests are shown in Fig. 2, by contrast (affricate, tone) and language (Cantonese, Mandarin). ANOVAs by participants and by items were conducted with correct response proportions as dependent variable and Group (adoptees, non-adopted Dutch controls), Test point (pretest, posttest), and Contrast (affricates, tones) as independent variables.

As Fig. 2 shows, the pattern of perception results was clear and consistent: the adoptee and control groups were virtually indistinguishable at the onset of the study, showing highly similar percentages of correct responses, but the adoptees were more likely to give correct responses in the posttest, i.e. at the end of the 10 blocks of training.

Main analyses for Cantonese revealed a significant interaction between Group and Test point, F_1 (1, 40) = 9.043, $p < .01$, $\eta^2_p = 0.184$, F_2 (1,30) = 4.399, $p < .05$, $\eta^2_p = 0.128$. Follow-up analyses further exploring this interaction showed a significant Group difference at posttest, F_1 (1, 40) = 5.691, $p < .05$, $\eta^2_p = 0.122$, F_2 (1, 30) = 8.597, $p < .01$, $\eta^2_p = 0.223$, but not at pretest (both $Fs < 1$). Further, there was no effect of Test point for either the adoptee group or the non-adopted Dutch control group (all $ps > 0.05$). Also, the main analyses showed a significant effect of Contrast (with numerically higher scores on Affricates than Tones) only in the F_1 analysis (F_1 (1, 40) = 5.257, $p < .05$, $\eta^2_p = 0.116$), but not in the F_2 analysis ($F_2 < 1$).

Separate analyses for Affricates and Tones revealed a complete lack of significant difference between the groups at the pretest (all $Fs < 1$). At posttest, in contrast, the adoptees statistically significantly outperformed the non-adopted Dutch controls on distinguishing the Affricate contrast, F_1 (1, 41) = 5.396, $p < .05$, $\eta^2_p = 0.116$, F_2 (1, 15) = 8.563,

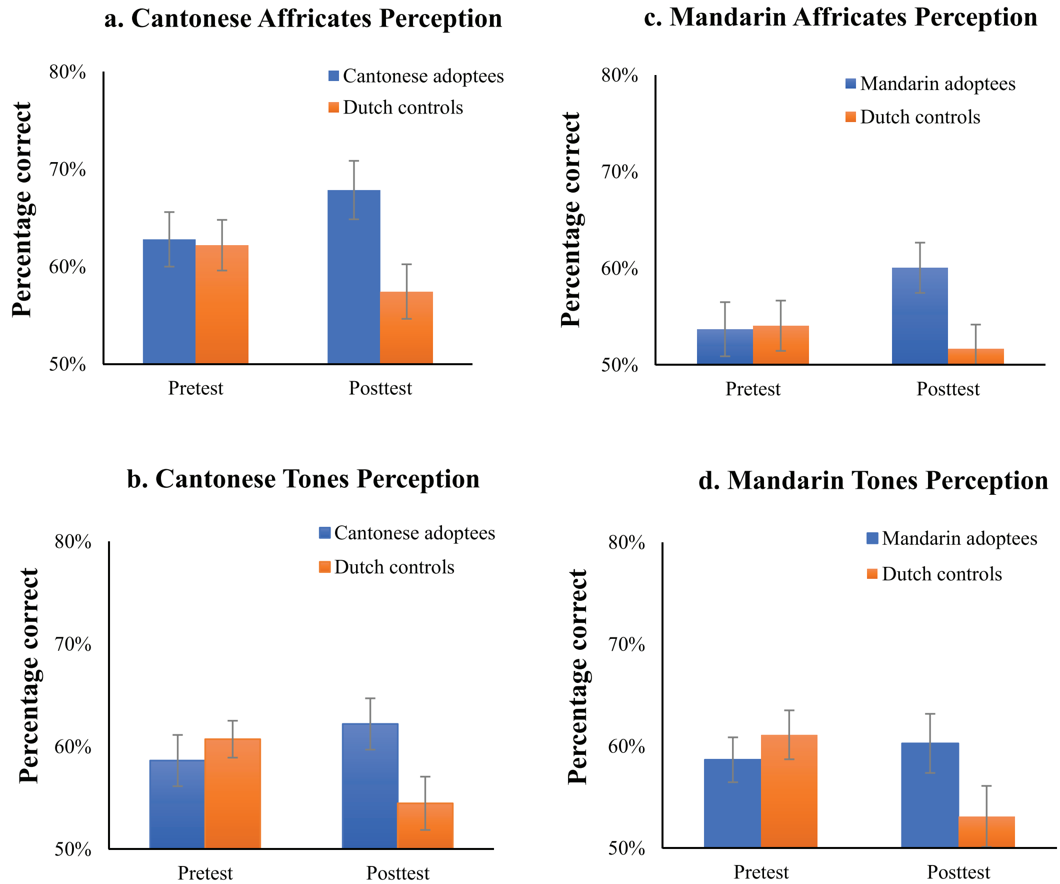

Fig. 2. Percentage of correctly perceived contrasts in pre- versus post-tests for the adoptee groups and their Dutch control groups, separately by language and contrast type; (a) Cantonese, affricates; (b) Cantonese, tones; (c) Mandarin, affricates; (d) Mandarin, tones. Error bars represent standard errors here and in all following figures.

$p < .05$, $\eta^2_p = 0.363$, while a similar trend did not reach statistical significance for the Tone contrast, F_1 (1, 41) = 3.492, $p = .069$, $\eta^2_p = 0.078$, F_2 (1, 15) = 2.012, $p > .05$, $\eta^2_p = 0.118$.

The results for Mandarin were virtually identical to those for Cantonese. Main analyses for Mandarin also showed a significant interaction between Group and Test point, F_1 (1, 46) = 4.361, $p < .05$, $\eta^2_p = 0.087$, F_2 (1, 30) = 9.051, $p < .01$, $\eta^2_p = 0.232$. Follow-up analyses again showed a significant Group difference at posttest, F_1 (1, 47) = 6.145 $p < .05$, $\eta^2_p = 0.116$, F_2 (1, 30) = 13.723, $p < .001$, $\eta^2_p = 0.314$, but not at pretest ($Fs < 1$). There was no effect of Test point for either the adoptee group or the Dutch control group (all $ps > 0.05$), and the main analyses showed no significant effect of or interaction with Contrast (all $ps > 0.05$).

Separate analyses for Affricates and Tones again confirmed that there were no significant differences between the groups at the pretest (all $Fs < 1$). At posttest, the adoptees outperformed the non-adopted Dutch controls statistically significantly for the Affricate contrast, F_1 (1, 48) = 5.499, $p < .05$, $\eta^2_p = 0.103$, F_2 (1, 15) = 7.133, $p < .05$, $\eta^2_p = 0.332$. Again there was a similar but non-significant trend for the Tone contrast, F_1 (1, 47) = 3.047, $p = .087$, $\eta^2_p = 0.061$, F_2 (1, 15) = 6.602, $p < .05$, $\eta^2_p = 0.306$.

2.4.3. Production

Separate analyses were carried out on the identification responses and on the goodness ratings produced by the Chinese adult native listeners, for all syllables produced by all adoptee and Dutch non-adopted control participants. As with the perception test analyses, responses with RTs more than two *SDs* from a native listener's mean RT were discarded from the analyses. For the Cantonese data set, 232 affricate identification responses (2.8%) and 345 tone identification responses (4.1%) were excluded, along with 73 affricate rating responses (1%) and 216 tone rating responses (2.9%). For the Mandarin data set, 369 responses (4.2%) were excluded from the affricate identifications and 380 (4.3%) from the tone identifications; 76 affricate rating responses (0.9%) were excluded, and 220 tone rating responses (2.5%). In all ANOVAs described below, Mauchly's Test of Sphericity was used to test whether the assumption of sphericity was violated, and if so, degrees of freedom were corrected using Greenhouse-Geisser estimates of Sphericity. The results for each language on each assessment are displayed in Figs. 3 and 4.

For both identification and rating, the proportions of correct responses were submitted to separate ANOVAs for affricates and tones with the independent variables Group (adoptees, non-adopted Dutch controls), Test point (pretest, posttest), and Target sound (Affricates: [ts] vs. [ts^h] in Cantonese and [tʂ] vs. [tʂ^h] in Mandarin; Tones: Tone 2 vs. Tone 5 in Cantonese and Tone 2 vs. Tone 3 in Mandarin). Given that our procedure necessarily involved independent sources of irrelevant participant-based variance, we carried out two separate ANOVAs, a first (F_{1A}) in which speakers were the random variable, and a second (F_{1B}) in which the native listeners were the random variable. Effects were only considered significant if the associated *p*-value was <0.05 in both analyses.

2.4.3.1. Identification. Fig. 3 shows the identification results; it is clear that adoptees' productions were more often correctly identified than the non-adopted Dutch controls' productions, and that this was the case for pretest as well as for posttest productions.

The Cantonese adoptees' pronunciations of both affricate and tone contrasts received significantly higher scores than those of the Dutch controls (Affricates [Fig. 3a], F_{1A} (1, 41) = 18. 201, $p < .001$, η^2_p =

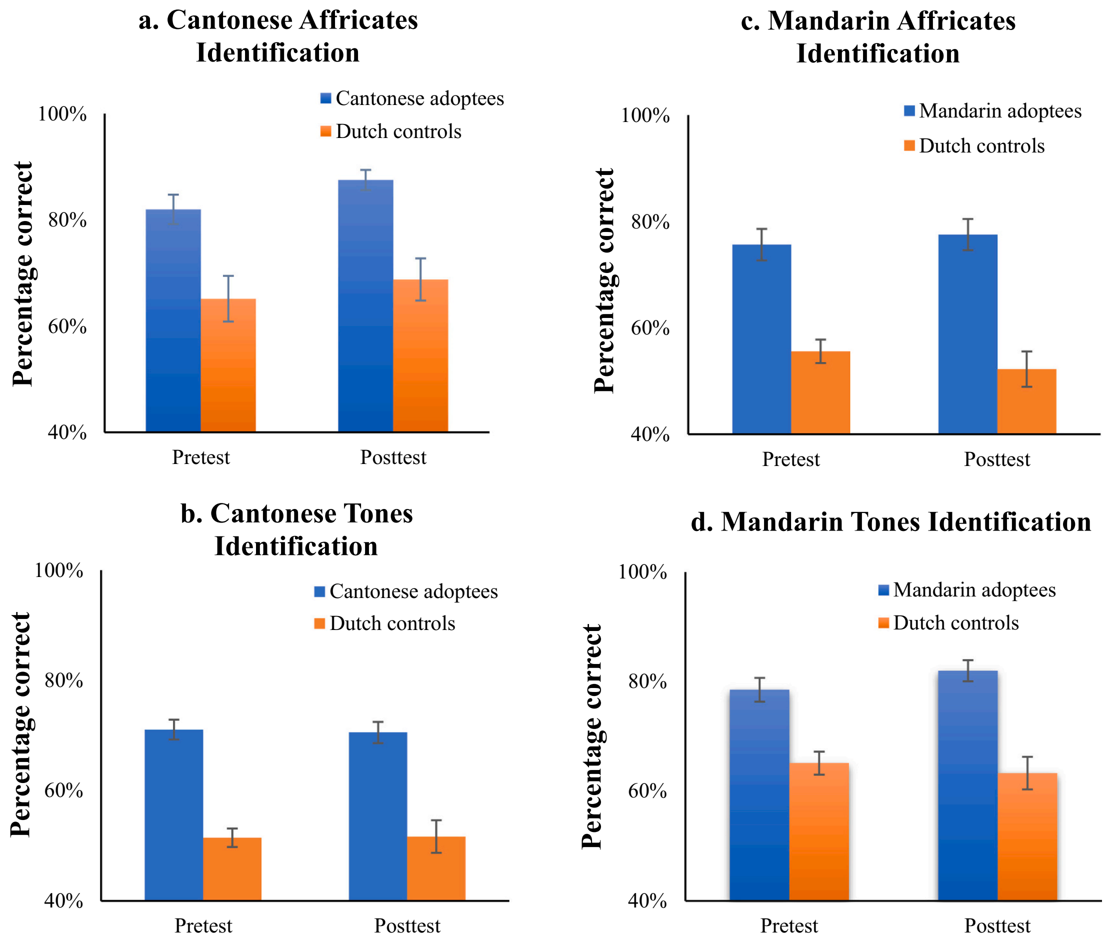

Fig. 3a-d. Percentage of correctly identified recordings uttered at pre- versus at post-tests by the adoptee groups and their Dutch control groups, separately by language and contrast type; (a) Cantonese, affricates; (b) Cantonese, tones; (c) Mandarin, affricates; (d) Mandarin, tones.

0.307, F_{1B} (1,23) = 835.927, $p < .001$, $\eta^2_p = 0.973$; Tones [Fig. 3b], F_{1A} (1, 41) = 53.095, $p < .001$, $\eta^2_p = 0.564$, F_{1B} (1, 23) = 314.269, $p < .001$, $\eta^2_p = 0.932$). Further, there was a significant effect of Target sound in the Cantonese tone identity block, with tone 2 being identified more accurately than tone 5 across the productions of all participants (Fig. 3e), F_{1A} (1, 41) = 13.647, $p < .001$, $\eta^2_p = 0.250$, F_{1B} (1, 23) = 11.781, $p < .01$, $\eta^2_p = 0.339$. There was no significant effect of Target sound in the Affricates (in the analysis by speakers $p > .05$; in the analysis by native listeners: F_{1B} (1, 23) = 10.511, $p < .01$, $\eta^2_p = 0.314$). The performance in neither group of children changed over time, shown by the absence of significant main effects of or interactions with Test point for either Affricates or Tones (all *ps* > 0.05 except for a main effect of Test point in the analysis by native listeners for Affricates: F_{1B} (1, 23) = 50.183, $p < .001$, $\eta^2_p = 0.686$).

In the Mandarin data set, adoptees again outperformed their non-adopted Dutch controls in producing the affricate contrast, reflected by a significant effect of Group (Fig. 3c), F_{1A} (1, 48) = 50.809, $p < .001$, $\eta^2_p = 0.514$, F_{1B} (1,21) = 285.865, $p < .001$, $\eta^2_p = 0.932$. There was a significant interaction between Test point and Target sound, F_{1A} (1, 48) = 4.264, $p < .05$, $\eta^2_p = 0.082$, F_{1B} (1, 21) = 59.225, $p < .001$, $\eta^2_p = 0.738$; follow-up analyses revealed that the Mandarin aspirated affricate [tʂʰ] was identified more accurately than the unaspirated affricate [tʂ] in both test points (Fig. 3f): Pretest, F_{1A} (1, 48) =25.489, $p < .001$, $\eta^2_p = 0.347$, F_{1B} (1, 21) = 74.772, $p < .001$, $\eta^2_p = 0.781$; posttest, F_{1A} (1, 48) =12.627, $p < .01$, $\eta^2_p = 0.208$, F_{1B} (1, 21) = 31.651, $p < .001$, $\eta^2_p = 0.601$. No effect of Test point was found, and no interaction between Test point and Group (all *ps* > 0.05), suggesting that the performance of neither group changed significantly over time.

For the Mandarin tones, a less general but still noticeable difference between the two participant groups appeared. Group and Target sound interacted significantly, F_{1A} (1, 48) = 8.848, $p < .01$, $\eta^2_p = 0.156$, F_{1B} (1, 21) = 95.734, $p < .001$, $\eta^2_p = 0.820$. Follow-up analyses on the interaction showed a significant Group difference only for Mandarin Tone 3 (Fig. 3g), with the adoptees' productions being identified significantly better than the controls', F_{1A} (1, 48) = 30.687, $p < .001$, $\eta^2_p = 0.390$, F_{1B} (1, 21) = 244.439, $p < .001$, $\eta^2_p = 0.921$, not for Mandarin Tone 2 (in the analysis by speakers, $p > .05$; in the analysis by native listeners: F_{1B} (1, 21) = 33.265, $p < .001$, $\eta^2_p = 0.613$). Additionally, there was a significant effect of Target sound, with Mandarin Tone 2 being significantly better identified than Mandarin Tone 3 (Fig. 3h) in productions by both the adoptees: F_{1A} (1, 24) = 16.074, $p < .01$, $\eta^2_p = 0.401$, F_{1B} (1, 21) = 13.534, $p < .01$, $\eta^2_p = 0.392$, and the Dutch controls: F_{1A} (1, 24) = 33.525, $p < .001$, $\eta^2_p = 0.583$, F_{1B} (1, 21) = 71.044, $p < .001$, $\eta^2_p = 0.772$. Finally, the main analysis showed no effect of or interactions with Test point (all *ps* > 0.05 except in the analysis by native listeners: main effect of Test point: F_{1B} (1, 21) = 18.451, $p < .001$, $\eta^2_p = 0.468$, interaction between Test point and Group: F_{1B} (1, 21) = 22.067, $p < .001$, $\eta^2_p = 0.512$), showing that neither the adoptees' nor the controls' performance improved significantly over time.

2.4.3.2. Rating. The results of the rating test are similar to those of the identification test. Fig. 4a and b show a clear group difference between the Cantonese adoptees and the non-adopted Dutch controls, with the adoptees being rated higher than the controls: Affricates, F_{1A} (1, 41) =

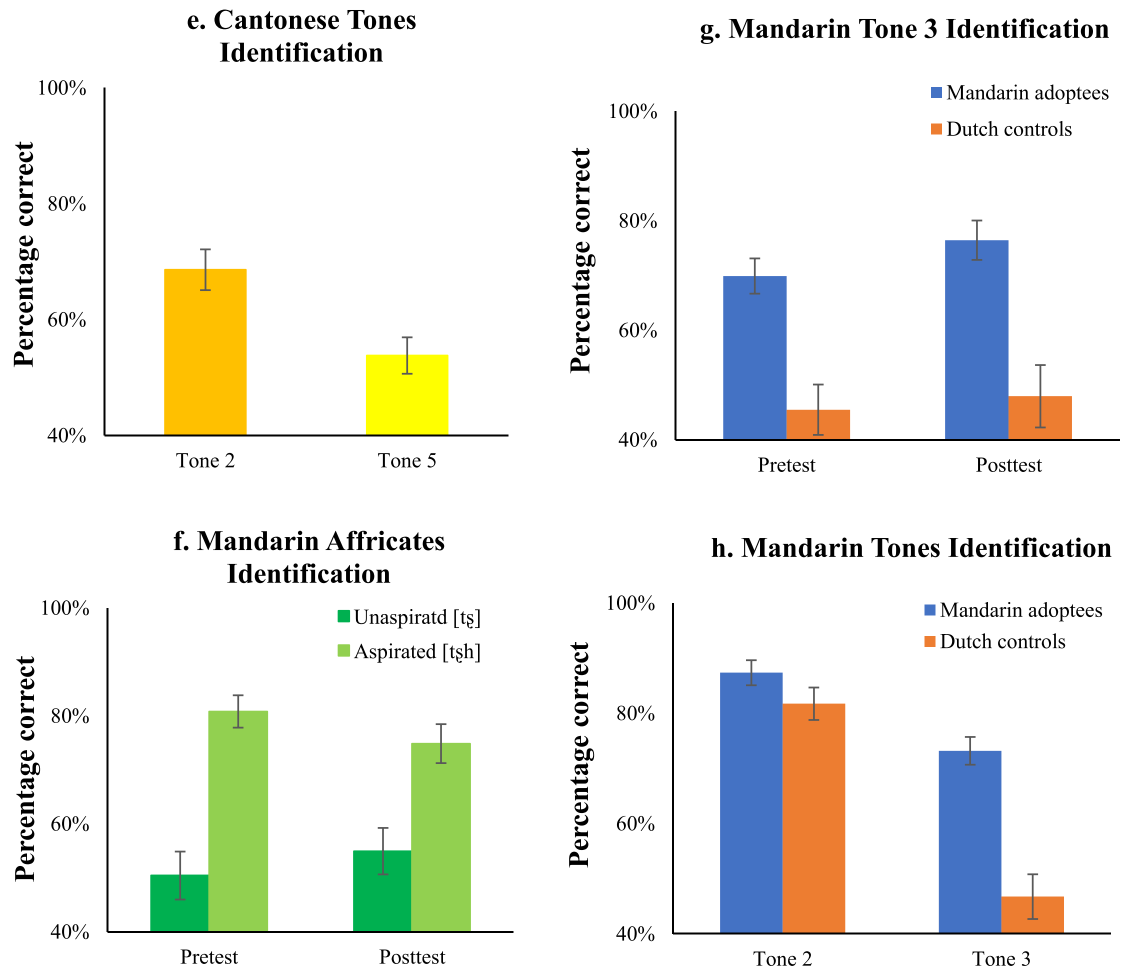

Fig. 3e-h. (e) Percentage of correctly identified recordings of Cantonese tone 2 versus tone 5, uttered by the adoptees and Dutch controls. (f) Percentage of correctly identified recordings of the Mandarin retroflex affricates unaspirated [tʂ] versus aspirated [tʂh], uttered at pre- versus at posttest by the adoptees and Dutch controls. (g) Percentage of correctly identified recordings of Mandarin tone 3, uttered at pre- versus at posttest by the adoptees versus Dutch controls. (h) Percentage of correctly identified recordings of Mandarin tone 2 versus tone 3, uttered at pre- versus at posttest by the adoptees versus Dutch controls.

12.414, $p < .01$, $\eta^2_p = 0.232$, F_{1B} (1,20) =529.828, $p < .001$, $\eta^2_p = 0.964$; Tones, F_{1A} (1, 41) =18.357, $p < .001$, η^2_p =. 309, F_{1B} (1, 20) = 33.115, $p < .001$, $\eta^2_p = 0.623$. Additionally, there was a significant effect of Target sound in the Cantonese tone rating block, showing that the production of Cantonese Tone 2 was rated significantly higher than Tone 5 (Fig. 4e), F_{1A} (1, 41) = 53.727, $p < .001$, $\eta^2_p = 0.565$, F_{1B} (1, 20) = 28.032, $p < .001$, $\eta^2_p = 0.584$, but no significant effect of Target sound in the affricates rating (both *F*s < 1). Finally, there was also a significant effect of Test point (Fig. 4f), F_{1A} (1, 41) = 6.629, $p < .05$, $\eta^2_p = 0.139$, F_{1B} (1, 20) = 20.978, $p < .001$, $\eta^2_p = 0.512$, showing that for both groups the ratings of their production of Cantonese tones improved over time, but the ratings of their affricate productions did not improve (by speakers $p > .05$, by native listeners F_{1B} (1, 20) = 25.414, $p < .001$, $\eta^2_p = 0.560$).

For Mandarin, the results of the rating test also strongly resemble those of the identification test. As shown in Fig. 4c, there was a significant Group effect in the Mandarin affricate rating block, with the Mandarin adoptees' productions being rated significantly higher than those of the Dutch controls, F_{1A} (1, 48) = 98.017, $p < .001$, $\eta^2_p = 0.671$, F_{1B} (1, 21) = 335.894, $p < .001$, $\eta^2_p = 0.941$. In addition, there was a significant effect of Target sound (Fig. 4g), with the aspirated affricate [$tʂ^h$] being produced significantly more acceptably than the unaspirated affricate [tʂ], F_{1A} (1, 48) = 5.514, $p < .05$, $\eta^2_p = 0.103$, F_{1B} (1, 21) = 15.349, $p < .01$, $\eta^2_p = 0.422$. There were no significant effects of or interactions with Test point (in the analysis by speakers, all *ps* > 0.05; by native listeners: main effect of Test point: F_{1B} (1, 21) = 18.687, $p < .001$, $\eta^2_p = 0.471$, interaction between Test point and Group: F_{1B} (1, 21) = 29.869, $p < .001$, $\eta^2_p = 0.587$), indicating participants did not improve their pronunciations of Mandarin affricate contrast over time.

As in Mandarin tone identification, Mandarin tone rating results also show a significant interaction between Group and Target sound, F_{1A} (1, 48) = 14.355, $p < .001$, $\eta^2_p = 0.230$, F_{1B} (1, 21) = 93.701, $p < .001$, $\eta^2_p = 0.817$. Follow-up analyses on the interaction showed that there was a significant Group difference (Fig. 4h) – with the adoptees being scored significantly higher than the controls – for both Mandarin Tone 2, F_{1A} (1, 48) = 9.63, $p < .01$, $\eta^2_p = 0.167$, F_{1B} (1, 21) =80.237, $p < .001$, $\eta^2_p = 0.793$, and Tone 3, F_{1A} (1, 48) = 30.597, $p < .001$, $\eta^2_p = 0.389$, F_{1B} (1, 21) = 154.726, $p < .001$, $\eta^2_p = 0.880$. Further, there was a significant effect of Target sound – with the production of Mandarin Tone 2 being rated significantly better than that of Tone 3 (Fig. 4h) – for both the Mandarin adoptees: F_{1A} (1, 24) = 70.114, $p < .001$, $\eta^2_p = 0.745$, F_{1B} (1, 21) = 99. 027, $p < .001$, $\eta^2_p = 0.825$, and the Dutch controls: F_{1A} (1, 24) = 114.636, $p < .001$, $\eta^2_p = 0.827$, F_{1B} (1, 21) = 143. 972, $p < .001$, $\eta^2_p = 0.873$. There was no significant effect of or interaction with Test point (all *ps* > 0.05 except for the main effect of Test point in the analysis by native listeners: F_{1B} (1, 21) = 35.475, $p < .001$, $\eta^2_p = 0.628$), showing no improvement of Mandarin tones over time.

To sum up, the production data show a consistent advantage in performance by the Chinese adoptees in comparison to the non-adopted Dutch controls, from the beginning. Except for the Cantonese tone rating, no changes over time (test point) among participants were found. Recall that this study involved no explicit training of production.

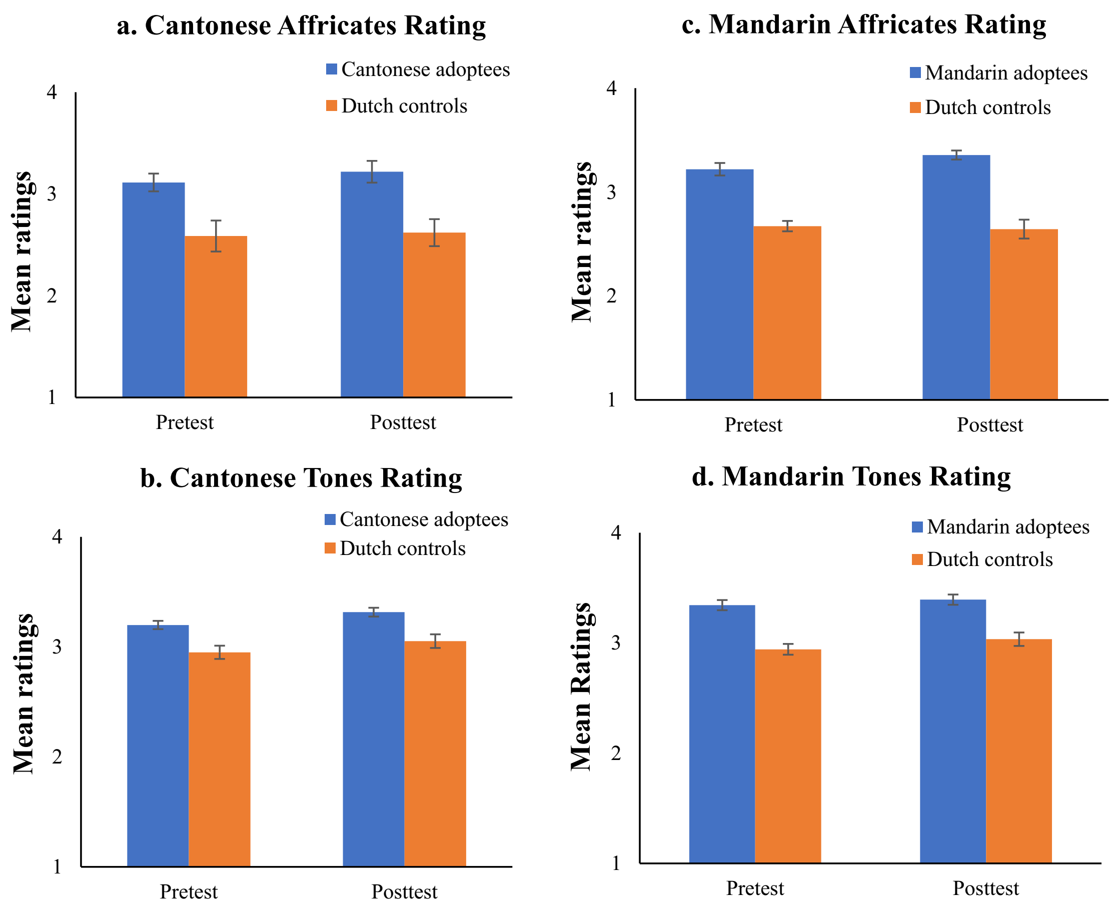

Fig. 4a-d. Mean ratings for the recordings uttered at pre- versus at posttest by the adoptee groups and their Dutch control groups, separately by language and contrast type; (a) Cantonese, affricates; (b) Cantonese, tones; (c) Mandarin, affricates; (d) Mandarin, tones. Ratings: 1: "completely different", 2: "a little similar", 3: "very similar", 4: "identical".

3. General discussion

While international adoptees learn their new language rapidly and efficiently, and appear to lose their birth language equally definitively, all is actually not lost. In this study of adoptees with an average age of seven, we predicted, based on prior studies with older language users, that the child participants would have stored abstract knowledge about the phonology of their birth language. This would then be available both to facilitate the learning process when our participants were asked to perceive sound contrasts of their birth language, and further to support successful speech production, such that the adoptees' productions were identified more correctly and rated more highly than the productions by control participants. Without the training exposure to the birth language, however, we predicted that they would present as dominant in the environmental language of their adopted country, and would have effectively no indication of any residual command of their birth language.

Our results confirmed most of these predictions, but interestingly, not all of them. The adoptee participants in our study had certainly, on the surface, lost command of their birth language. Their vocabulary knowledge was essentially nil. Their perceptual identification of both segmental and tonal contrasts in their birth language was, without training, no better than that of their matched Dutch-speaking control participants. The abstract residual knowledge was definitely in place, enabling the adoptees to improve across the 10 blocks of training at identifying the birth language contrasts – significantly more so than the non-adopted Dutch controls could. All this was in line with our expectations. Further, although all Cantonese schoolchildren are bilingual, and the Cantonese adoptees are likely to have received at least some exposure to Mandarin pre-adoption (e.g. via media or at school), the patterns of results were the same for the two language groups.

In several ways, though, our findings went beyond our predictions. For one, we found that the residual knowledge that adoptees had at their disposal was not limited to segmental categories but also included tones. Note that the prior literature on adult adoptees' knowledge retention has largely relied on distinctions of segmental phonology. Some test cases have concerned contrasts on a dimension that does not feature at all in the adopted language. The Korean three-way distinction based on articulatory dynamics, that Choi, Broersma, and Cutler (2017) and Choi, Cutler, and Broersma (2017) found to be rapidly relearned by Korean-born adoptees in The Netherlands, is of that kind; in an articulatory space that in the adopted language has only the single voiceless plosive sound [t], Korean as birth language offers three kinds of [t], and likewise three kinds of [k] or [p]. Other test cases have concerned unfamiliar differences in features such as place of articulation (Singh et al., 2011: dental vs retroflex) or manner of articulation (Bowers et al., 2009: plosive vs implosive), and these too have proved learnable in adulthood. Our segmental results show that the same learning benefit, for a segmental contrast involving aspiration of affricates, can be observed in children who are between one and ten years post adoption. We further added evidence that contrasts that are relatively harder to acquire in childhood (the Cantonese case in comparison to the Mandarin case) are also relatively harder to acquire even in short-term learning tasks such as in our study.

For tone contrasts, empirical evidence on adoptees' preservation of their experience exists. In the brains of adoptees whose birth language is

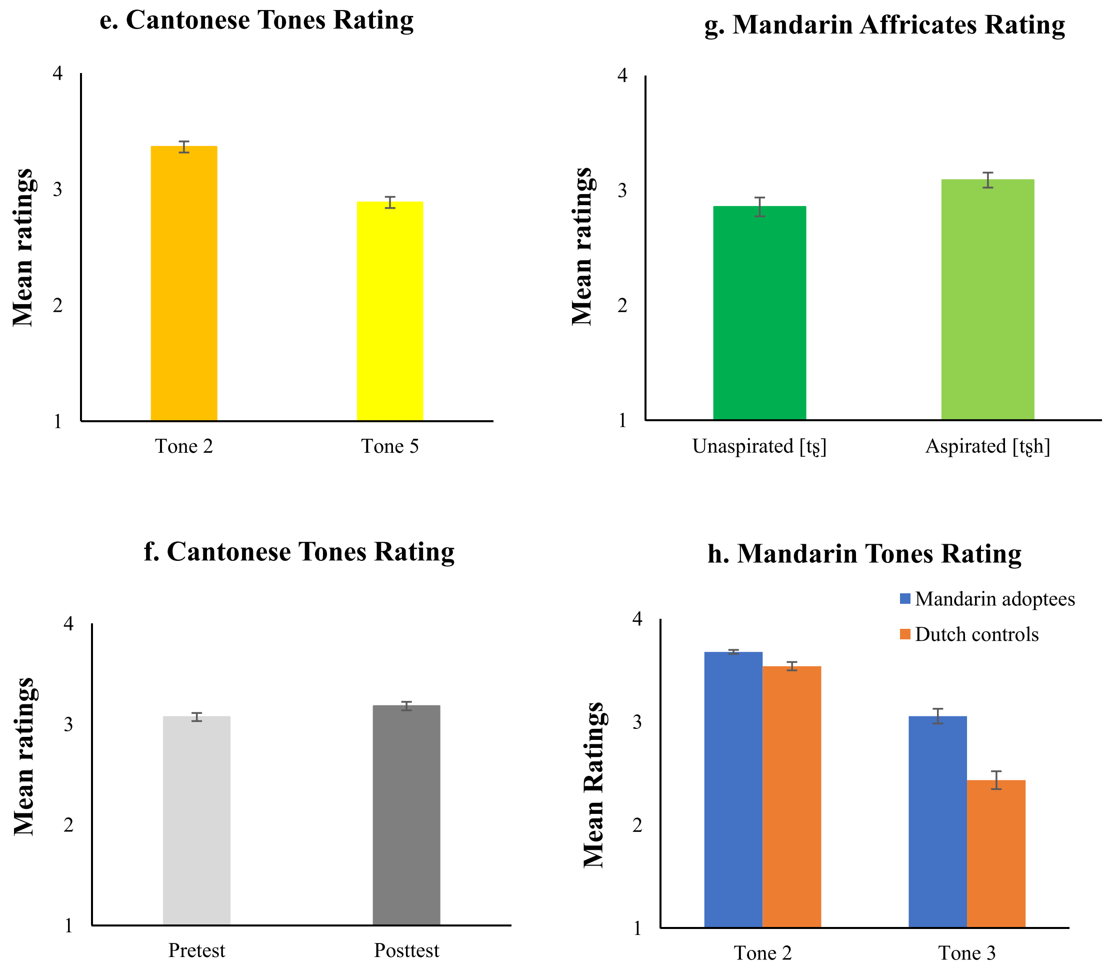

Fig. 4e-h. (e) Mean ratings for Cantonese tone 2 versus tone 5, uttered by the adoptees and Dutch controls. (f) Mean ratings for Cantonese tones uttered at pretest versus at posttest by the adoptees and Dutch controls. (g) Mean ratings for Mandarin retroflex affricates unaspirated [tʂ] versus aspirated [tʂh], uttered by the adoptees and Dutch controls. (h) Mean ratings for Mandarin tone 2 and tone 3, uttered by the adoptees versus Dutch controls.

a tone language, while their adopted language is not, tones are processed with the cortical language areas, just as is the case in the brains of native users of that birth language. Control participants' brains, in contrast, treat the tones as non-linguistic (Pierce et al., 2014, 2015). In our perception results, tone contrasts showed somewhat less evidence of very rapid learning in this short testing period than the segmental contrasts did. This, we assume, has its source in the nature of tonal versus segmental distinctions; segmental contrasts are typically defined in terms of binary features, while tones allow more gradient expression even within the features of high, medium, low, and falling versus rising. The learning of tone categories and their category boundaries in a tone language is an uneven progression from early discrimination through later failure to discriminate to a relatively late attainment of native categorization skill (Götz et al., 2018). The learning of tone realizations thus appears to be more demanding than the learning of segmental contrasts. The adoptee-control patterns observed for tones and segments were nonetheless similar.

Most strikingly, however, we found a clear difference between our perception and production results. The perception findings provided the predicted evidence for what happens when adoptees are trained on a difficult contrast not relevant to the adoptive language currently in use: their retained abstract knowledge from their early language facilitates rapid perceptual improvement. Thus, the perceptual pretest showed the adoptees to be performing no better than the non-adopted Dutch controls; the perceptual posttest, however, showed that even in this short training period the adoptee group had learned enough to outperform the control group, and in the case of segmental contrasts significantly so. This is the pattern that has been observed with adults who had had early exposure to a subsequently lost language (Bowers et al., 2009; Singh et al., 2011; Choi, Broersma, & Cutler, 2017, Choi, Cutler, & Broersma, 2017). Our findings therefore confirm that such knowledge is accessed by young children in perceptual learning in the same way as by adults.

The production measures in our study revealed a different pattern: the adoptees' productions were more accurate than the non-adopted Dutch controls' productions, and this remained the same from pre- to posttest. In the light of the perception findings, this is surprising: accurate imitation is possible even when perception (ability to identify) cannot support it. The literature on adoptees' birth language knowledge offers no ready explanation for this finding, nor do our other results. Some potential explanations can however be readily discounted.

First, the good imitation skills of the adoptees reflected in the good identification and rating scores cannot be ascribed exclusively to these participants' age; i.e., it is not enough to assume that children under 10 years tend to be accurate mimics. If this were so, then the Dutch controls, who were matched on age to the adoptees, should perform as well as the adoptees; but they did not manage such accuracy at all. Precision in production is the adoptees' accomplishment only.

Second, the relative familiarity of the contrasts from the birth language, against a background of the adopted language, also offers no explanation. Dutch does not have lexical tone or anything phonologically similar to it, whereas it does have affricates (although not these contrasts involving aspiration). Effects of contrast type familiarity would predict differing production success for the segmental versus the tonal contrasts, but no such difference was observed; the adoptees' productions in both cases were identified more accurately and rated more highly for pronunciation goodness than were the productions by the control children.

Third, the relative difficulty of the contrasts does not offer an explanation; there were differences in difficulty of the individual contrasts, known from the developmental literature and duly observed in our data as well, but these differences did not systematically interact with the main effect of Group, as an explanatory contribution would require. Only the Mandarin tone contrast showed an interaction with Group. This contrast, it will be recalled, pits a simple rise (Tone 2) against a more complex tone (Tone 3), with the simpler tone target having unsurprisingly an advantage in Mandarin children's acquisition. As can be seen from the perception data in Fig. 2d, Mandarin Tone 2 was the single contrast case in which the non-adopted Dutch controls produced slightly (albeit not significantly) higher scores in the identification pretest. This suggests that Tone 2 was initially quite easy for the Dutch controls. A simple rise can occur in Dutch question intonation, and this might have made it easier to match in the pretest, for both controls and adoptees. However a single contrast condition where the adoptees had no advantage in the perception pretest clearly brings us not a step nearer to explaining the general advantage for the adoptees at both pre- and posttest of production.

The Mandarin Tone 2 case does, however, underline the difference that we have already pointed out between our perception and production tasks. The perception task required categorization; not just same-different discrimination of two adjacent tokens, but storage of the first heard token followed by comparison of the two subsequent tokens to that first representation. The production task, which was not trained, was simple imitation of a single token. It was, accordingly, easier to perform (and particularly so in the case of Mandarin Tone 2 which was probably the production target closest to any Dutch equivalent). The relative success in the two tasks is thus not surprising; it has long been known from studies of L2 processing by adults that simple imitation performance will usually outstrip identification performance (Flege & Eefting, 1987, 1988; Hao & de Jong, 2016; Sheldon & Strange, 1982).

Indeed, the imitation of speech is a well-studied task in adult language performance. It has been shown that successful imitation of spoken language can draw on abstract knowledge of the imitation target (Hao & de Jong, 2016; Llompart & Reinisch, 2018; Mitterer & Ernestus, 2008; Nielsen, 2011). Sato et al. (2013) have argued that imitation is however dependent on *perceptual* learning; a perceptual advantage will predict a production advantage. This is fully consistent with the findings from adults who had once been child adoptees; Choi, Cutler, and Broersma (2017) found that production accuracy and goodness after training were predicted by rapidity of perceptual learning from training.

The child adoptees in the present study were certainly in possession of residual abstract knowledge about their birth language; this was obvious from their improvement across the training in the identification task (compared to the lack of improvement by the non-adopted Dutch controls). Contrary to the performance of the adult adoptees in Choi, Cutler, and Broersma (2017), where production scores were correlated with perception scores, the child adoptees' performance in the present study showed no such identification-production dependency. There was also no improvement across time in the production task; the adoptees' imitation was good from the outset, and the controls' imitation seems to have been as good as it was going to get. It is thus unlikely that our child adoptees' good imitation skills at pretest were due to more rapid accessibility of the same abstract knowledge that enabled their improvement on the identification task; if there had been such access, it should presumably have improved early performance on the identification task as well.

Instead, we view this result in terms of the type of knowledge involved in the two tasks. In identification, it is necessary to know where the boundaries between the segmental or tonal categories fall, in order to be able to assign the input token correctly. In imitation, explicit categorization may help if it is available (Flege & Eefting, 1988), but imitation can also occur without it (Hao & de Jong, 2016; Llisterri, 1995; Llompart & Reinisch, 2018; Sheldon & Strange, 1982). Our findings are consistent with a scenario in which category boundary knowledge, and other category-defining information such as the relative weighting of phonetic cues, is available to these adoptees only once explicit training has awakened the traces of their stored earlier linguistic experience. Other simpler knowledge may nevertheless still be available, and this can include awareness of dimensions that are not used in the adopted language, but in prior experience were relevant for distinguishing between words. Contrasts signaled by fundamental frequency can be one such case. Another can be particular properties that play a part in phonemic category contrasts (e.g., aspiration that can distinguish consonantal sounds). Enough of this type of knowledge still available, even if not directly drawn upon during the five years or so of their new life, could provide the adoptees with a crucial advantage in the imitation task.

With contrasts such as these that have no counterpart in the adopted language because they involve different phonological dimensions, it is also possible that gestural memory for how to realise contrasts in such dimensions is still available from the time of birth-language exposure, not necessarily coupled to stored memory of the actual birth-language phonology. This would be an interesting finding in itself and could be tested by comparing adoptees' and controls' performance on learning, in this case, tone contrasts or contrasts involving aspiration that are similar but not identical to those used by the birth language.

We suggest that our child adoptees thus did, in a small way, draw on their birth language experience in the imitation task; to outperform the non-adopted Dutch control participants, it was enough that they knew what general properties of the signal to imitate, while the controls did not have such awareness. (Had the adoptees been separated from their birth language for six times as long, as Choi, Broersma, and Cutler (2017) and Choi, Cutler, and Broersma (2017) adoptees were, it seems likely that such general information about language characteristics and the dimensions of phonological contrast would no longer be available). To do better than the controls in the perception task, however, the adoptees needed to learn how to identify the relevant category boundary signals. For this, it really helped them that they could access stored knowledge from that earlier experience (knowledge that either contains or can facilitate the computation of detailed phonological contrasts; knowledge that also, as the adult results suggest, is likely to remain accessible to them even for decades to come). Even then, the process of (re)learning the relevant cues took some time.

The present study thus makes three contributions to our knowledge of birth-language effects in language processing by international adoptees. First, the relearning benefit of adoptees' stored birth-language knowledge for perception is confirmed, and it is extended to the case of child adoptees, who can profit from this benefit early in their new life, just as do teenagers and adults later on. Second, our study has also confirmed that production of unfamiliar contrasts, at least in an imitation task, may succeed independently of perception. Third, this latter finding indicates in turn that some retained experience of general properties of the birth language, potentially gestural as well as categorical, may be still available to international adoptees in childhood, even though that language has been to all intents and purposes forgotten by these child users, and even though no similar effects of general birth-language structure seem to remain available for long enough that they also appear in productions by adults.

Acknowledgements

This research was supported by the Max Planck Institute for Psycholinguistics and by the Donders Institute for Brain, Cognition and Behaviour (Centre for Cognition) at Radboud University, both situated in Nijmegen, The Netherlands. Additional support is acknowledged from NWO-Spinoza (AC), and, during manuscript preparation, from NWO-Vidi (MB), and the ARC Centre of Excellence for the Dynamics of Language (AC). The authors

declare no competing interests. Ethics compliance to the following standards was ensured: Radboud University Code of Academic Integrity (https://www.ru.nl/english/research/other-research/academic-integrity/#Appendix1Violationsofintegrityinscienceandscholarship), the Association of Universities in the Netherlands (VSNU) Code of Conduct on Scientific Practice (http://www.vsnu.nl/files/documents/Netherlands%20Code%20of%20Conduct%20for%20Research%20Integrity%202018.pdf), the ALLEA European Code of Conduct for Research Integrity (https://allea.org/code-of-conduct/) and the World Medical Association Declaration of Helsinki (https://www.wma.net/publications/wma-doh-1964-2014/).

As the above paragraph illustrates, academic life brings people together from afar, often to form the lasting collaborations that are so vital for science and its successful performance. The present authors were born in China, The Netherlands and Australia, respectively, and have been collaborating for more than a decade. Earlier, the third author also collaborated for more than a decade with the (Argentinian-raised) honoree of this special issue. As many other contributors to the special issue will attest, among the wealth of reasons for honoring Jacques is that he was a true master of that collaborative art.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104788.

References

Antoniou, M. (2019). The advantages of bilingualism debate. *Annual Review of Linguistics, 5*, 395–415.

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese phonology. 102*. Walter de Gruyter.

Bowers, J. S., Mattys, S. L., & Gage, S. H. (2009). Preserved implicit knowledge of a forgotten childhood language. *Psychological Science, 20*, 1064–1069.

Burns, T. C., Yoshida, K., Hill, K., & Werker, J. (2007). Bilingual and monolingual infant phonetic development. *Applied Psycholinguistics, 28*, 455–474.

Choi, J., Broersma, M., & Cutler, A. (2015). Enhanced processing of a lost language: Linguistic knowledge or linguistic skill?. In *Interspeech 2015: 16th annual conference of the International Speech Communication Association* (pp. 3110–3114).

Choi, J., Broersma, M., & Cutler, A. (2017). Early phonology revealed by international adoptees' birth language retention. *Proceedings of the National Academy of Sciences, 114*, 7307–7312.

Choi, J., Broersma, M., & Cutler, A. (2018). Phonetic learning is not enhanced by sequential exposure to more than one language. *Linguistic Research, 35*, 567–581. https://doi.org/10.17250/khisli.35.3.201812.006.

Choi, J., Cutler, A., & Broersma, M. (2017). Early development of abstract language knowledge: Evidence from perception–production transfer of birth-language memory. *Royal Society Open Science, 4*, 160660.

Duanmu, S. (2007). *The phonology of standard Chinese*. OUP Oxford.

Flege, J. E., & Eefting, W. (1987). Production and perception of English stops by native Spanish speakers. *Journal of Phonetics, 15*, 67–83.

Flege, J. E., & Eefting, W. (1988). Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation. *The Journal of the Acoustical Society of America, 83*, 729–740.

Gervain, J., Nespor, M., Mazuka, R., Horie, R., & Mehler, J. (2008). Bootstrapping word order in prelexical infants: A Japanese–Italian cross-linguistic study. *Cognitive Psychology, 57*, 56–74.

Goriot, C., McQueen, J. M., Unsworth, S., Van Hout, R., & Broersma, M. (2020). Perception of English phonetic contrasts by Dutch children: How bilingual are early-English pupils? *PLoS One, 15*(3), Article e0229902. https://doi.org/10.1371/journal.pone.0229902.

Götz, A., Yeung, H. H., Krasotkina, A., Schwarzer, G., & Höhle, B. (2018). Perceptual reorganization of lexical tones: Effects of age and experimental procedure. *Frontiers in Psychology, 9*, 477.

Gussenhoven, C. (1992). Dutch. *Journal of the International Phonetic Association, 22*, 45–47.

Hao, M., Shu, H., Xing, A., & Li, P. (2008). Early vocabulary inventory for Mandarin Chinese. *Behavior Research Methods, 40*, 728–733.

Hao, Y. C., & de Jong, K. (2016). Imitation of second language sounds in relation to L2 perception and production. *Journal of Phonetics, 54*, 151–168.

Hua, Z., & Dodd, B. (2000). The phonological acquisition of Putonghua (Modern standard Chinese). *Journal of Child Language, 27*, 3–42.

Isurin, L. (2000). Deserted island or a child's first language forgetting. *Bilingualism: Language and Cognition, 3*, 151–166.

Johnson, E. K. (2016). Constructing a proto-lexicon: An integrative view of infant language development. *Annual Review of Linguistics, 2*, 391–412.

Kaushanskaya, M., & Marian, V. (2009). Bilingualism reduces native-language interference during novel-word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 829.

Kovács, Á. M. (2009). Early bilingualism enhances mechanisms of false-belief reasoning. *Developmental Science, 12*, 48–54.

Kovács, Á. M., & Mehler, J. (2009). Flexible learning of multiple speech structures in bilingual infants. *Science, 325*, 611–612.

Krakow, R., & Roberts, J. (2003). Acquisition of English vocabulary by young Chinese adoptees. *Journal of Multilingual Communication Disorders, 1*, 169–176.

Krakow, R., Tao, S., & Roberts, J. (2005). Adoption age effects on English language acquisition: Infants and toddlers from China. *Seminars in Speech and Language, 26*, 33–43.

Levi, S. (2018). Another bilingual advantage? Perception of talker-voice information. *Bilingualism: Language and Cognition, 21*, 523–536.

Llisterri, J. (1995). Relationships between speech production and speech perception in a second language. *Proceedings of the XIIIth International Congress of Phonetic Sciences, 4*, 92–99.

Llompart, M., & Reinisch, E. (2018). Imitation in a second language relies on phonological categories but does not reflect the productive usage of difficult sound contrasts. *Language and Speech, 62*, 594–622, 0023830918803978.

Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition, 109*, 168–173.

Nicoladis, E., & Grabois, H. (2002). Learning English and losing Chinese: A case study of a child adopted from China. *International Journal of Bilingualism, 6*, 441–454. https://doi.org/10.1177/13670069020060040401.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics, 39*, 132–142.

Pierce, L. J., Chen, J. K., Delcenserie, A., Genesee, F., & Klein, D. (2015). Past experience shapes ongoing neural patterns for language. *Nature Communications, 6*, 10073.

Pierce, L. J., Klein, D., Chen, J. K., Delcenserie, A., & Genesee, F. (2014). Mapping the unconscious maintenance of a lost first language. *Proceedings of the National Academy of Sciences, 111*, 17314–17319.

Ramon-Casas, M., Swingley, D., Sebastián-Gallés, N., & Bosch, L. (2009). Vowel categorization during word recognition in bilingual toddlers. *Cognitive Psychology, 59*, 96–121.

Roberts, J., Krakow, R., & Pollock, K. (2003). Three perspectives on language development of children adopted from China. *Journal of Multilingual Communication Disorders, 1*, 163–168.

Roberts, J. A., Pollock, K. E., Krakow, R., Price, J., Fulmer, K. C., & Wang, P. P. (2005). Language development in preschool-age children adopted from China. *Journal of Speech, Language, and Hearing Research, 48*, 93–107.

Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, J. L., & Nguyen, N. (2013). Converging toward a common speech code: Imitative and perceptuo-motor recalibration processes in speech production. *Frontiers in Psychology, 4*, 422.

Scott, K. A., Pollock, K., Roberts, J. A., & Krakow, R. (2013). Phonological processing skills of children adopted internationally. *American Journal of Speech-Language Pathology, 22*, 673–683.

Scott, K. A., Roberts, J. A., & Krakow, R. (2008). Oral and written language development of children adopted from China. *American Journal of Speech-Language Pathology, 17*, 150–160.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics, 3*, 243–261.

Singh, L., Liederman, J., Mierzejewski, R., & Barnes, J. (2011). Rapid reacquisition of native phoneme contrasts after disuse: You do not always lose what you do not use. *Developmental Science, 14*, 949–959.

Snedeker, J., Geren, J., & Shafto, C. L. (2007). Starting over: International adoption as a natural experiment in language development. *Psychological Science, 18*, 79–87. https://doi.org/10.1111/j.1467-9280.2007.01852.x.

So, L. K. H., & Dodd, B. J. (1995). The acquisition of phonology by Cantonese-speaking children. *Journal of Child Language, 22*, 473–496. https://doi.org/10.1017/S0305000900009922.

Stepanov, A., Pavlič, M., Stateva, P., & Reboul, A. (2018). Children's early bilingualism and musical training influence prosodic discrimination of sentences in an unknown language. *The Journal of the Acoustical Society of America, 143*, EL1–EL7.

Tremblay, M. C., & Sabourin, L. (2012). Comparing behavioral discrimination and learning abilities in monolinguals, bilinguals and multilinguals. *The Journal of the Acoustical Society of America, 132*, 3465–3474.

Tsao, F. M. (2008). The effect of acoustical similarity on lexical-tone perception of one-year-old Mandarin-learning infants. *Chinese Journal of Psychology, 50*, 111–124.

Ventureyra, V. A. G., Pallier, C., & Yoo, H.-Y. (2004). The loss of first language phonetic perception in adopted Koreans. *Journal of Neurolinguistics, 17*, 79–91. https://doi.org/10.1016/S0911-6044(03)00053-8.

Werker, J. F. (1986). The effect of multilingualism on phonetic perceptual flexibility. *Applied Psycholinguistics, 7*, 141–155.

Werker, J. F., & Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics, 37*, 35–44. https://doi.org/10.3758/BF03207136.

Zhou, W. (2015). *Assessing birth language memory in young adoptees* (unpublished doctoral dissertation). The Netherlands: Max Planck Institute for Psycholinguistics https://pure.mpg.de/rest/items/item_2166297_10/component/file_2273014/content.

Zhu, H. (2002). *Phonological development in specific contexts: Studies of Chinese-speaking children. 3*. Multilingual Matters.

Zink, I., & Lejaegere, M. (2002). *N-CDIs: Lijsten voor communicatieve ontwikkeling [NCDIs: Lists for communicative development]*. Leuven, Belgium: Acco.