**Children’s identification of unfamiliar voices on both target-present and target-absent lineups**

**Calderwood, L., 1\* McKay, D. R.,1 and Stevenage, S.V.2**

1 Department of Psychology, University of the West of Scotland

2 Department of Psychology, University of Southampton

Correspondence:

Correspondence should be addressed to Dr Lesley Calderwood at:

School of Media, Culture and Society, University of the West of Scotland, Paisley Scotland, PA1 2BE

Email: [Lesley.Calderwood@uws.ac.uk](mailto:Lesley.Calderwood@uws.ac.uk); Tel: 0141 848 3895

**Children’s identification of unfamiliar voices on both target-present and target-absent lineups**

**Abstract**

A robust finding from the eyewitness literature is that children are as accurate as adults on target-present lineups from the age of five years, whereas they continue to make an erroneous false positive identification from a target-absent lineup up until around fourteen years of age (Pozzulo & Lindsay, 1998). The current study explores whether the same pattern occurs when voices are used instead of faces. A total of 334 participants from six age groups (6-7-year-olds, 8-9-year-olds, 10-11-year-olds, 12-13-year-olds, 14-15-year-olds and adults) listened to a 30 second audio clip of an unfamiliar voice and were then presented with either a six person target-present or target-absent voice lineup. Overall, participants were more accurate with target-present than target-absent lineups. Moreover, performance on target-present lineups showed adult-like levels of attainment by 8-9 years of age. In contrast, performance on target-absent lineups was extremely poor, with all age groups tending to make a false identification. Confidence was higher when participants made correct rather than incorrect decisions for both types of lineup and this did not change with increasing age. Given these results, both child and adult earwitness evidence needs to be treated with considerable caution.

# Introduction

In most criminal cases the perpetrator has been seen by a victim or eyewitness(es) and their visual descriptions are used to aid the identification process. However, there are often cases in which the victim’s or witness’s memory of the perpetrator’s voice can provide a useful clue to identification. There may even be occasions when the voice is the only clue to identification such as when crimes are committed in the dark, over the phone, or when the perpetrator is wearing a disguise or visibility is reduced. Whilst we know a lot about the reliability (and fallibility) of eyewitness identification, very little is known about the reliability of earwitness identification and even less is known about children’s voice identification abilities. In particular, there is a lack of understanding of children’s voice identification abilities when the target is absent from the test set. The purpose of the present paper is to examine children’s voice identification abilities with a specific emphasis on performance in target-absent testing conditions.

Research with adults has consistently shown that voice identification is much poorer than face identification (see Yarmey, 1995, and Stevenage & Neil, 2014, for reviews). For example, Öhman, Eriksson and Granhag, (2012), tested adult recognition of an unfamiliar voice using a seven voice lineup after a two week delay. Surprisingly, only 19% of adults could correctly identify the target voice. Although this result was significantly above chance, performance was surprisingly poor when compared to unfamiliar face recognition. This difficulty recognising voices is evident even when examining familiar voice recognition (Damjanovic & Hanley, 2007; Hanley & Damjanovic, 2009; Hanley, Smith & Hadfield, 1998; Yarmey, Yarmey, Yarmey, & Parliament, 2001), leading to the proposal of a weaker route for voice identification than for face recognition (Stevenage, Hugill & Lewis, 2012).

Against this backdrop, there is little research as yet on children’s recognition of voices. Consequently, we do not know whether they too show relatively poor performance with voices compared to faces. Early work, however, suggests that, like adults, children’s recognition of *familiar* voices, is far from perfect. For example, when presented with a 4 second clip of a familiar cartoon voice and asked to point to the corresponding cartoon picture, children as young as 3 years performed at only 61%, with performance rising to around 81-86% by 4-5 years (Spence, Rollins & Jerger, 2002). Addressing concerns over lack of realism, lack of explicit identification, and potential unfamiliarity with the cartoon voices, Murry and Cort, (1971), asked 20 9-10-year-olds to identify the voices of 10 personally familiar classmates. Accuracy of identification reached 47% when listeners were presented with a 5 second vowel sound, but reached 95% accuracy when listeners heard a sentence or a paragraph of speech. However, these performance levels may have been artificially inflated by the use of personally familiar stimuli drawn from a small set.

In terms of unfamiliar voice processing, results have suggested this to be a particularly challenging task for children (Mann, Diamond & Carey, 1979). Indeed, when given a relatively taxing task involving the presentation of a target, and then the selection of a matching voice from two alternatives, performance was below chance level in 6 and 8 year olds, and rose only to around 60% by about 10 years of age. Moreover, when given a slightly easier version of the task, involving a same/different judgement with two voice clips presented sequentially, performance was only marginally above chance in the youngest group (6 year olds) and rose only to around 70% by about 10 years of age.

Taking these results together, two observations are clear: (i) adult levels of voice recognition performance are relatively poor compared to face recognition performance in equivalent designs, and (ii) children show similarly poor performance. Importantly, however, Mann et al.’s (1979) work explored unfamiliar voice recognition only through 2AFC target-present trials and, as such, it offered no insight into children’s performance when the original speaker was absent at test. This is an important oversight, given evidence that children may find target-absent testing conditions much harder than target-present conditions. Indeed, in the face domain, whilst adult-like levels of performance are attained by children as young as 5-6 years old when the target is present, children as old as 14 years continue to make false positive selections from the lineup when the target is absent (see Pozzulo & Lindsay, 1998, for a review). As such, current evidence may indeed overestimate the performance of children when recognising voices.

In a recent review of this area, it is suggested that children’s high levels of false identifications in target-absent face lineups may be driven by a social pressure to pick someone from the lineup in order to ‘please the experimenter’ rather than due to immature face processing abilities (Havard, 2014). Such an explanation bears a resemblance to children’s performance in Piagetian conservation tasks, where children who are repeatedly asked the same question may think that they should give a different answer (Rose & Blank, 1974). In support of this view, it has been found that where children can select a ‘mystery man’ or a silhouette from a target-absent lineup, false identifications are greatly reduced (Havard & Memon, 2013; Zajac & Karageorge, 2009). Moreover, Dunlevy and Cherryman, (2013), included a tree in target-present and target-absent lineups and told child participants that they should select the tree if they thought the culprit was ‘not there and was hiding behind the tree’. The provision of this ‘tree’ option reduced false positive identifications dramatically in 6-7-year-old children and is consistent with the theory that children’s high false positive rates may reflect a lack of complete understanding of the task, and an implicit desire to choose.

In order to test this social pressure explanation of target-absent mistakes, children’s performance is tested here with target-present and target-absent voice lineups. As such, the current study provides a necessary re-examination of children’s voice processing abilities, whilst also incorporating the critical target-absent testing condition. In this regard, there is only one study which provides a point of reference (Öhman, Eriksson & Granhag, 2011). The authors tested 7-9-year-olds, 11-13-year-olds and adults, all of whom listened to a 40 second voice clip with a single target speaker before making an identification decision from a 7-person lineup two weeks later. Importantly, half of the participants were given a target-present lineup and the remainder were given a target-absent lineup. The results were curious in several regards. First, the data indicated better performance in 11-13-year-olds than in younger and adult groups in both target-present and target-absent conditions, and this pattern sits in opposition to evidence for a developmental dip at this age both in face recognition (Carey, Diamond & Woods, 1980) and voice recognition (Mann et al., 1979). Furthermore, the data suggested better performance in target-absent than in target-present conditions across all ages despite high levels of false identifications in the target-absent case. The use of a single target voice, together with a 2 week delay, suggested that Öhman et al.’s test may have been particularly difficult, and suggest that a replication of these findings would be beneficial.

***The Present Study***

Based on the previous review, the present study will employ both target-present and target-absent trials in a voice lineup paradigm. In extension to Öhman et al.’s (2011) work, two target voices will be used so that results may be generalised beyond the single target voice used in previous studies. In addition, the difficulty of the task will be reduced through repeated exposure to test voices, and immediate rather than delayed testing, in order to avoid potential floor effects. On the basis of previous work, it is expected that performance will improve with age on both target-present and target-absent lineups, reaching adult levels by about 10 years of age. More importantly though, the present design enables examination of performance in target-absent conditions in order to provide a convergent test of the social pressure hypothesis of false positive identifications.

**Method**

***Design***

A 6 x 2 between-groups design was used in which voice identification was investigated across six age groups (6-7-year-olds, 8-9-year-olds, 10-11-year-olds, 12-13-year-olds, 14-15-year-olds, and adults) on both target-present and target-absent lineups. Accuracy and confidence were the dependent variables.

***Participants***

A total of 334 participants took part in the present study, with 163 participants receiving a target-present lineup and 171 receiving a target-absent lineup. All were drawn from the West of Scotland area and thus were familiar with the accent of the speakers whilst being unfamiliar with the speakers themselves. This project was approved by the School of Media, Culture & Society Ethics Committee at the University of the West of Scotland. In accordance with this approval, written parental consent was obtained for all child participants. Verbal assent was also obtained on the day of testing from all child participants. Adult participants provided their own written consent to take part in the study. Participants were randomly allocated to the lineup condition with similar numbers in each age group as detailed in Table 1.

Table 1 shows participant characteristics for each age group

|  |  |  |
| --- | --- | --- |
| **Age group** | **Mean Age** | **Gender** |
| 6-7-year-olds (N= 44) | 6 years, 8 months  Range (6, 0 – 7,8) | 24 F, 20 M |
| 8-9-year-olds (N= 54) | 9 years, 0 months  Range (8,0 – 9, 9) | 26 F, 28 M |
| 10-11-year-olds (N= 56) | 10 years, 7 months  Range (10, 0 – 11, 9) | 30 F, 26 M |
| 12-13-year-olds (N= 70) | 12 years, 5 months  Range (12, 0 – 13,3) | 36 F, 34 M |
| 14-15-year-olds (N= 62) | 14 years, 7 months  Range 14,0 – 15, 3 | 33 F, 29 M |
| Adults (N= 48) | 28.84 years, SD = 7.97  Range (18-54 years) | 25 F, 23 M |

***Materials***

Two speech clips were obtained from a total of 24 Scottish female speakers, aged between 25-35 years, recruited from the West of Scotland area. All had a standard accent typical of the region and were free from speech impediments. The voice clips contained free rather than scripted speech. To obtain this, participants were shown a picture of two scenes, a farmyard and a fairground, and were asked to describe them. From these recordings, a thirty second clip of one scene was selected for the study phase and an eight second clip of the other scene was selected for the lineup phase. In this way, the content differed between the study and lineup phases and so could not be used to help with the identification of the target speaker.

From this database of 24 speakers, the ratings from nine local participants were used to select target speakers, target-replacements (for target-absent lineups), and the foils. Two target speakers were selected on the basis that they had no distinguishing characteristics such as pitch, speaking rate and modularity. The remaining 22 speakers were compared to the targets and the most similar voice for each target (as rated on a 5-point scale) was selected as the target-replacement for the target-absent lineups. The next five most similar voices for each target were selected as the foils for both the target-present and target-absent lineups. Analysis of the ratings confirmed that there was no significant difference in the perceived similarity of the two target replacement voices used in the target-absent lineups to their respective target voice (target 1 similarity = 3.11, target 2 similarity = 3.22, *t*(8) = 0.32, *p* > 0.05). The remaining voices were not used in the experiment.

Audacity 1.3 Beta was used to edit the clips of the targets, the foils and the target- replacements to produce a thirty second clip of the two targets describing the photograph of the farm for the study phase, and an eight second clip of each speaker describing the fairground for the lineup phase. These timings were in line with previous work using a similar paradigm (e.g., Stevenage, Clarke & McNeill, 2012; Öhman et al., 2012). For each of the recordings, audio quality was 705 kbits per second, 44.1 Hz, 16 bit, and volume was equivalent across clips.

From these stimuli, four lineups were created: a target-present and a target-absent lineup for Speaker 1 and a target-present and a target-absent lineup for Speaker 2. Each participant heard one lineup and care was taken to vary the positioning of the target, or the target-replacement in each lineup for each participant.

All voices were presented through a PowerPoint presentation which participants played at their own pace. Testing took place in a quiet environment, however, headphones were used to minimise distraction from ambient noise.

***Procedure***

Both child and adult participants were tested in small groups of four or five within a quiet area of their school or university. Each participant completed the lineup using a laptop and a set of headphones. They were instructed to listen carefully to the target voice and encouraged to focus on the voice rather than the content in readiness for a subsequent recognition test. Each target voice was then played for thirty seconds. Following this, participants were told that they would hear six short clips and they would be asked to consider whether any of them sounded similar to the target. They were warned that the target may not be one of the six voices and that half of the lineups were target-present and the other half were target-absent. After listening to each of the six clips, participants were then told that they would hear the six clips again and that this time they had to make a decision for each voice about whether it was the same person as the target voice. For each voice, they marked either Yes or No on a sheet of paper beside them. The lineup task was completed either when all six test voices had been heard, or when a Yes decision had been indicated. Participants were then asked to rate how confident they felt about their decision on a 5 point scale, with 1 indicating a guess and 5 indicating that they were very sure.

**Results**

Preliminary analysis examined performance with target voice as a factor. There were no significant main effects or interactions so all analyses reported here are collapsed across both target voices.

***Accuracy***

In terms of accuracy, participants who listened to a target-present lineup could respond with a hit (correct identification of the target), false identification (selection of a foil) or a miss (incorrectly saying that the target was not there). Similarly, participants who listened to a target-absent lineup could respond with a correct rejection (correctly saying that the target was not there) or a false identification (selecting the target replacement or any one of the foils). The pattern of performance for each age group is presented and analysed separately for target-present and target-absent lineups in order to determine whether there are different patterns of responding for each type of lineup.

*Target-present lineups*

Figure 1 shows the percentage of responses for each possible response, for each age group. As shown in Figure 1, 6-7-year-olds were more likely to make a false identification, whereas all other age groups were more likely to make a correct identification. ‘Miss’ responses (rejecting the lineup and saying that the target was not present) were uncommon in the two older groups and the adult group, while all other groups (especially the 10-11-year-olds) showed a fair number of ‘miss’ responses.

Figure 1shows percentages for each possible response as a function of age in the target-present lineups

A chi-square test was used to explore the association between age (6-7, 8-9, 10-11, 12-13, 14-15-year-olds and adults) and response type (correct identification, false identification and miss). Given that there were several cells with an expected count less than 5, Fisher’s Exact test was used. This confirmed a significant association between age and response (*p* < 0.001). Analysis of the standardised residuals indicated that this was due to a particularly low level of performance in the 6-7-year-olds, together with a particularly good level of performance (as indicated by very few ‘miss’ decisions) from 12-13-year-olds. This is suggestive of Mann et al.’s, (1979), demonstration of the attainment of adult levels of performance by the age of 10 years, followed by a developmental dip and then recovery of performance levels thereafter. However, the adult level of attainment was observed by 8-9 years of age rather than by 10 years. Indeed, when the 6-year-olds were removed from the analysis, Fisher’s Exact test showed that there was no significant association between age and response, *p* = 0.22, This suggests that adult levels of performance were attained by 8-9 years of age and that there were no further improvements in performance.

*Target-absent lineups*

As can be seen in Figure 2, the rate of false identifications in target-absent lineups was high and remained high across all age groups.

Figure 2 shows percentages for each possible response as a function of age in the target-absent lineups

As above, a chi-square test was used to explore the association between age (6-7, 8-9, 10-11, 12-13, 14-15-year-olds and adults) and response type (correct rejection, false identification). This revealed no significant association between the two variables, χ²(5, *N* = 171) = 3.37, *p* = 0.65*.* In fact, all age groups were more likely to make a false identification than correctly report that the target was absent.

Overall, the results suggested some improvement in performance with target-present lineups, when 6-7-year-olds were compared to older participants. The absence of any association between age and response after this age group suggested that adult levels of performance had been attained by 8-9 years of age. In contrast, no age-related improvements emerged with target-absent lineups. In fact, performance was rather poor in target-absent lineups compared to target-present lineups through the inappropriate tendency to select a voice rather than indicate that the target was not present.

***Confidence***

Once participants had made their decision, they were asked to rate how confident they were in their decision on a scale from 1-5 (where 1 indicated that they were guessing and 5 indicated that they were very confident that they were correct).

*Target-present lineups*

The mean confidence scores for correct and incorrect responses on target-present lineups are presented below in Figure 3. Confidence tended to be higher for most age groups when responses were correct, however, the 6-7-year-olds and the 14-15-year-olds were slightly more confident when incorrect.

Figure 3 shows mean confidence scores (and standard errors) across target-present lineups in each age group

A 6 (age group) x 2 (lineup accuracy: correct vs incorrect) between-groups Analysis of Variance (ANOVA) was conducted on the confidence data. This revealed a main effect of lineup accuracy only, with confidence being higher when correct (M = 4.11, CI = 3.96 - 4.26) than when incorrect (M = 3.63, CI = 3.31 - 3.96), *F*(1, 149) = 5.01, *p* = 0.027, . There was no significant main effect of age, *F*(5, 149) = 0.699, *p* = 0.625, and no significant interaction between age and accuracy, *F*(5, 149) = 0.96, *p* = 0.44, suggesting that the improvement in accuracy with age (as noted above) was not mirrored by a similar increase in confidence.

*Target-absent lineups*

The mean confidence scores for correct and incorrect responses on target-absent lineups are presented below in Figure 4. Confidence was higher for all age groups when responses were correct.

Figure 4 shows mean confidence scores (and standard errors) across target-absent lineups in each age group

For target-absent lineups, a 6 x 2 between groups ANOVA was again conducted. The results mirrored those with target-present lineups in all respects. Specifically, a significant main effect of lineup accuracy emerged, with higher confidence when correct (M = 3.98, CI = 3.66 - 4.29) than when incorrect (M = 3.57, CI = 3.38 - 3.76), *F*(1, 151) = 4.71, *p* = 0.032, 0.03. As before, there was no significant main effect of age, *F*(5, 151) = 0.41, *p* = 0.84, and no significant interaction, *F*(5, 151) = 0.19, *p* = 0.97.

In summary, for both target-present and target-absent lineups, confidence was significantly higher when participants’ decisions were correct than when incorrect regardless of age group. In the target-absent case, stable confidence levels mirrored stable accuracy levels. Interestingly, however, in the target-present case, improvements in accuracy with age were not accompanied by an increase in confidence. This said, in both cases, the effect sizes were small and the interpretation of confidence data should therefore be treated cautiously.

To further explore the confidence-accuracy relationship we also calculated confidence-accuracy calibration curves which provide information about accuracy at each level of confidence, and an indication of overconfidence/underconfidence (Juslin, Olsson & Winman, 1996). This method is thought to provide a more informative representation of the confidence-accuracy relationship than traditional correlational methods (Mickes, 2015; Wixted & Mickes, 2012).

Figure 5 shows the confidence-accuracy calibration curves for both target-present and target-absent lineups. We collapsed across both identities for each type of lineup and across all age groups in order that there were enough participants in the analysis (upwards of 200 per condition, see for example, Weber & Brewer, 2004). The diagonal line shows where data points would fall if confidence and accuracy were perfectly calibrated. Points that fall above this line reflect underconfidence, whereas points falling below the line are indicative of overconfidence.

Figure 5 shows confidence-accuracy calibration curves for both target-present and target-absent lineups with standard error bars. The diagonal line shows perfect calibration.

Based on visual inspection, the participants' overall self-rated confidence seems to be reasonably well calibrated to their overall accuracy for target-present lineups with a tendency to be underconfident when correct, however, self-ratings of confidence do not calibrate well with overall accuracy for target-absent lineups. Participants appear to be overconfident on target-absent lineups. Thus, the calibration data do give a useful indication of accuracy on the different types of lineups, however, this is based on only one data point per participant, and caution is again urged when interpreting the confidence data.

**Discussion**

The aim of this study was to examine children’s voice matching performance with a view to testing the social pressure hypothesis of children’s high false positive responses, as found consistently in target-absent lineups in the eyewitness literature. We found that performance was more accurate on target-present than target-absent lineups when voices were used, a finding that echoes the pattern found with faces, and suggests that children find it very difficult to correctly reject a target-absent lineup regardless of the stimulus used. The particularly poor performance in the target-absent condition, relative to the target-present condition, was surprising given the opposite pattern in Öhman et al.’s (2011) study. It is also of particular applied interest given the real-world consequences of inappropriately selecting from a target-absent lineup. As noted earlier, it has been suggested that children’s high false positive identifications may be driven largely by a social pressure to choose and that false identifications dramatically reduce when a mystery man, a silhouette or a tree is included in the lineup, allowing children to identify when they think that the target is absent (see Havard, 2014, for a review). It would be interesting to see if the inclusion of a similar option in voice lineups would also see a similar reduction in false positive identifications in child responses.

In the current study, we found no developmental improvement on the target-absent lineups; adults also tended to make a false identification in the target-absent testing conditions. Similar difficulties on target-absent trials have been found with adults by several authors using both shorter (e.g., Stevenage, et al., 2012; Philippon, Cherryman, Bull & Vrij, 2007) and longer delays (e.g., Van Wallendael, Surace, Parsons, & Brown, 1994). Indeed, Van Wallendael et al. found that on target-absent trials, all but one participant (out of 76) made a false identification, regardless of the length of the delay. It seems to be the case that, unlike faces, false identifications for voices are high in childhood and remain high throughout adulthood. This finding has serious implications for those working with earwitnesses as there was very little delay in the current study between hearing the initial voice and making the lineup identification. Thus, the difficulty of voice recognition is emphasised for both children and adults through the use of these simplified testing conditions.

Several explanations may exist to account for the high rate of false identifications in target-absent trials for voices. For instance, it is possible that the lineups had been created such that the similarity between target and foil voices was too high to support effective target identification. Alternatively, the memory load associated with a necessarily sequential voice lineup created task demands that were too difficult. In both cases, however, it is difficult to understand why the performance in target-absent trials was notably worse than that in target-present trials where the same issues existed. A further possibility is that the sequential nature of the presentation of the voices influenced the number of misidentifications on the target-absent line-ups. Participants listened to each voice once and then listened to each voice again, and were required to respond Yes or No to each voice as opposed to listening to all of the voices twice and then making one single decision as recommended by the Home Office (2003). In future research, it would be useful to investigate whether children and adults continue to make a rate of false positive decisions using this more forensically valid technique. It is also worth considering the possibility of interference effects when recognising voices (Stevenage et al., 2013) in which the presentation of intervening voices between study and test can dramatically impair the memory for the original voice. This may have the capacity to account for poorer performance in target-absent trials than in target present trials, as the lack of the target means that every voice in the target absent lineup is a distractor voice.

***Development of Voice Matching Skills***

Another aim of the current study was to investigate the developmental pattern shown by Mann et al. (1979), using a different paradigm in order to extend the literature on the development of voice matching skills. In this regard, the results did not entirely support the previous pattern of development across the age range. Rather than showing attainment of adult levels of performance by 10 years of age, the analysis of target-present performance actually indicated adult levels of attainment in slightly younger participants of 8-9 years of age. There was also some evidence of a developmental dip in the 10-11-year-olds, with a significant recovery in performance by 12-13-year-olds, as shown by the absence of any ‘miss’ responses when the target was present in the lineup. This result sits well alongside that of Öhman et al. (2012) who showed no improvement in target-present voice performance when 11-13-year-olds were compared with adults. When examining performance on target-absent lineups, the lack of any improvement with age was clear. In fact, all participants showed equivalent and rather poor performance with target-absent lineups compared to target-present lineups and had a tendency to select a voice inappropriately from the lineup.

Together, these results sit at odds with those obtained from face lineups where both target-present performance (e.g., Bruce et al., 2000; Megreya & Bindemann, 2015) and target-absent performance (e.g., Pozzulo & Lindsay, 1998; Havard, 2014) have been shown to improve with age. This discrepancy is unlikely to be due to floor effects, as the use of a long (30 second) clip at study, and the use of an immediate test, ensured that performance exceeded that of previous studies (c.f., Öhman et al., 2012). Instead, the current results underline the difficulty of the voice matching task, immediate or otherwise, compared to a face matching task.

***A Reflection on Confidence***

In turning to a consideration of confidence ratings, the value of witness confidence ratings has been debated for many years. In the eyewitness literature, there traditionally appeared to be a weak positive relationship at best between confidence and accuracy. Confidence was generally not regarded as a reliable predictor of identification accuracy (see Sporer, Penrod, Read, & Cutler, 1995, for a review). A similar observation has been made regarding confidence in the earwitness literature, with early studies often reporting a low or non-significant relationship between confidence and voice identification accuracy in adults (Yarmey, 1995; Yarmey, 2001). Öhman et al. (2011), found no evidence for a relationship between confidence and accuracy for their child and adult participants and, whilst Öhman, Eriksson and Granhag, (2013), did not formally measure confidence, when they asked participants if they thought they would be able to recognise a target voice, 86% of children and 63% of adults responded positively. In reality, only 13% of children and 4% of adults could correctly identify the target voice indicating a high level of overconfidence in all age groups. Within the current study, confidence emerged as being significantly higher when participant decisions were correct than when incorrect, across both target-present and target-absent lineups. This pattern held regardless of participant age suggesting that participants knew when they were correct. This suggests that confidence could be a useful indicator of accuracy in earwitness studies. However, the data from the confidence-accuracy calibration indicates that when participants were correct on the target-present lineups, there was a tendency to be underconfident in their response with the opposite pattern occurring in target-absent lineups.

The current results suggest that confidence may be a useful indicator of earwitness accuracy, however, these effects were associated with very small effect sizes, and the calibration curves show different patterns for target-present and target-absent lineups. Therefore, caution should be encouraged when assessing the reliability of confident earwitnesses. In addition, the current study has been designed to maximise participants’ chances of accurately recalling the voice by using an immediate recognition test with no time delay. Due to these methodological testing conditions, confidence ratings may be over-inflated in our study compared to real life earwitnesses where there is a substantial delay between hearing the perpetrator’s voice and making an identification. In future studies, it would be useful to investigate whether children show the same pattern between accuracy and confidence using a more ecologically valid design with a substantial delay between hearing the target voice and completing the lineup task.

***Limitations and Future Work***

One point worth reflecting on within the current study was the fact that all voices were obtained from adult speakers. Given own-age effects when recognising faces (Bonner & Burton, 2004; Rhodes & Anastasi, 2012), this may represent a confound in the current design. More specifically, it may have been more appropriate to test each age-group with voices drawn from their own age. This said, the design of the present study, whilst perhaps not optimal, would have biased performance in favour of the adult participants. In this regard, the fact that all bar the youngest children performed at a level comparable to the adults here is perhaps notable. Nevertheless, future work would be well-directed to test voice recognition across the age range by using age-relevant voices.

***Conclusion***

In conclusion, the present results show that voice recognition, even on immediate testing, was a rather difficult task. Performance was better in target-present than target-absent lineups, possibly reflecting a tendency to make a positive selection from the lineup rather than reporting that the target was not there. In target-present lineups, adult levels of attainment appeared to be demonstrated by 8-9 years of age and recovered by 12-13 years of age after a slight developmental dip. Of particular concern was the very poor performance on target-absent lineup trials, with real-world implications being felt for the innocent police suspect who may be selected from the lineup inappropriately. Given these results, both child and adult earwitness evidence needs to be treated with considerable caution. Even if a witness may have confidence in their identification, the current data suggests that earwitness performance may be too poor to rely on in court.

**References**

Bonner, L., & Burton, A. M. (2004). 7–11‐year‐old children show an advantage for matching and recognizing the internal features of familiar faces: Evidence against a developmental shift. *The Quarterly Journal of Experimental Psychology Section A*, *57*(6), 1019-1029.

Bruce, V., Campbell, R. N., Doherty‐Sneddon, G., Langton, S., McAuley, S., & Wright, R. (2000). Testing face processing skills in children. *British Journal of Developmental Psychology*, *18*(3), 319-333.

Carey, S., Diamond, R., & Woods, B. (1980). Development of face recognition: A maturational component? *Developmental Psychology*, *16*(4), 257.

Damjanovic, L., & Hanley, J. R. (2007). Recalling episodic and semantic information about famous faces and voices. *Memory & Cognition*, *35*(6), 1205-1210.

Dunlevy, J. R., & Cherryman, J. (2013). Target-absent eyewitness identification line-ups: Why do children like to choose. *Psychiatry, Psychology and Law*, *20*(2), 284-293.

Hanley, J. R., & Damjanovic, L. (2009). It is more difficult to retrieve a familiar person's name and occupation from their voice than from their blurred face. *Memory*, *17*(8), 830-839.

Hanley, J. R., Smith, S. T., & Hadfield, J. (1998). I recognise you but I can't place you: An investigation of familiar-only experiences during tests of voice and face recognition. *The Quarterly Journal of Experimental Psychology: Section A*, *51*(1), 179-195.

Havard, C. (2014). Are children less reliable at making visual identifications than adults? A review. *Psychology, Crime & Law*, *20*(4), 372-388.

Havard, C., & Memon, A. (2013). The mystery man can help reduce false identification for child witnesses: Evidence from video line‐ups. *Applied Cognitive Psychology*, *27*(1), 50-59.

Home Office, UK. (2003). Advice on the Use of Voice Identification Parades, Circular 057/2003, Police Leadership and Powers Unit, London.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1304.

Mann, V. A., Diamond, R., & Carey, S. (1979). Development of voice recognition: Parallels with face recognition. *Journal of Experimental Child Psychology*, *27*(1), 153-165.

Megreya, A. M., & Bindemann, M. (2015). Developmental improvement and age-related decline in unfamiliar face matching. *Perception*, *44*(1), 5-22.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition*, *4*(2), 93-102.

Murry, T., & Cort, S. (1971). Aural identification of children's voices. *Journal of Auditory Research*, *11(3),* 260-262.

Öhman, L., Eriksson, A., & Granhag, P. A. (2011). Overhearing the planning of a crime: Do adults outperform children as earwitnesses? *Journal of Police and Criminal Psychology*, *26*(2), 118-127.

Öhman, L., Eriksson, A., & Granhag, P. A. (2012). Enhancing Adults' and Children’s Earwitness Memory: Examining Three Types of Interviews. *Psychiatry, Psychology and Law*, *20*(2), 216-229.

Öhman, L., Eriksson, A., & Granhag, P. A. (2013). Angry Voices from the Past and Present: Effects on Adults' and Children's Earwitness Memory. *Journal of Investigative Psychology and Offender Profiling*, *10*(1), 57-70.

Philippon, A. C., Cherryman, J., Bull, R., & Vrij, A. (2007). Earwitness identification performance: The effect of language, target, deliberate strategies and indirect measures. *Applied Cognitive Psychology*, *21*(4), 539-550.

Pozzulo, J. D., & Lindsay, R. C. L. (1998). Identification accuracy of children versus adults: a meta-analysis. *Law and Human Behavior, 22(5),* 549-570.

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: a meta-analytic and theoretical review. *Psychological Bulletin, 138(1),* 146-174.

Rose, S.A., & Blank, M. (1974). The potency of context in children's cognition: An illustration through conservation. Child Development. 45(2), 499–502.

Spence, M. J., Rollins, P. R., & Jerger, S. (2002). Children's recognition of cartoon voices. *Journal of Speech, Language, and Hearing Research*, *45*(1), 214-222.

Sporer, S. , Penrod, S., Read., J. & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin. 118*. 315-327.

Stevenage, S. V., Clarke, G., & McNeill, A. (2012). The “other-accent” effect in voice recognition. *Journal of Cognitive Psychology, 24,* 647-653.

**Stevenage, S. V., Hugill, A. R.,** & **Lewis, H. G.** (2012). Integrating voice recognition into models of person perception. Journal of Cognitive Psychology *24(4),* 409–419.

Stevenage, S.V., & Neil, G.J. (2014). Hearing Faces and Seeing Voices: The Integration and Interaction of Face and Voice Processing. *Psychologica Belgica. 54(3),* 266–281.

Stevenage, S. V., Neil, G. J., Barlow, J., Dyson, A., Eaton-Brown, C., & Parsons, B. (2013). The effect of distraction on face and voice recognition. *Psychological research*, *77*(2), 167-175.

Van Wallendael, L. R., Surace, A., Parsons, D. H., & Brown, M. (1994). ‘Earwitness’ voice recognition: Factors affecting accuracy and impact on jurors. *Applied Cognitive Psychology*, *8*(7), 661-677.

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied*, *10*(3), 156.

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science*, *7*(3), 275-278.

Yarmey, A. D. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law*, *1*(4), 792.

Yarmey, A. D. (2001). Expert testimony: Does eyewitness memory research have probative value for the courts? *Canadian Psychology*, *42*(2), 92-

Yarmey, A. D., Yarmey, A. L., Yarmey, M. J., & Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology*, *15*(3), 283-299.

Zajac, R., & Karageorge, A. (2009). The wildcard: A simple technique for improving children's target-absent lineup performance. *Applied Cognitive Psychology*, *23*(3), 358–368.