Received: 3 June 2023 | Revised: 31 July 2023 | Accepted: 3 September 2023

DOI: 10.1002/acp.4133

RESEARCH ARTICLE

WILEY

Wisdom of the inner crowd benefits both face and voice matching

Robin S. S. Kramer[1] | Emily Flanagan[1] | Alex L. Jones[2] | Georgina Gous[1]

[1]School of Psychology, University of Lincoln, Lincoln, UK

[2]Department of Psychology, Swansea University, Swansea, UK

Correspondence
Robin S. S. Kramer, School of Psychology, University of Lincoln, Lincoln LN6 7TS, UK.
Email: remarknibor@gmail.com

Abstract

Identification often involves determining whether two face photographs or voice samples originated from the same person. Here, we investigated the wisdom of the (outer) crowd (averaging two individuals' responses to the same trial) and inner crowd (averaging the same individual's responses to the same trial after completing the test twice) as routes to increased performance. Participants completed the same face (Experiment 1) or voice matching test (Experiment 2) twice with no delay. In addition, we reanalysed previously collected data where these tests were completed with a one-week interval between sessions. For both tests, whether with or without a delay, inner crowds outperformed participants' individual test responses and were equivalent to outer crowds of two participants. Taken together, we demonstrate the use of inner crowds as a robust method of improvement during identification. In contexts where outer crowds are not feasible, agencies should consider inner crowds as a promising alternative.

KEYWORDS

aggregation, face matching, voice matching, wisdom of the crowd, wisdom of the inner crowd

1 | INTRODUCTION

A number of real-world situations involve confirming an individual's identity using their facial appearance. For instance, one might identify a traveller using their passport or a suspect from CCTV footage. Typically, a comparison is made either between two different photographs or a single image and a 'live' presentation of the face. The accuracy of this 'face matching' process can have significant security implications and so researchers have been focussed on performance in this task, along with potential routes to improvement.

While our abilities to match and recognise familiar faces are at, or near, ceiling-level performance (Bruce et al., 2001; Burton et al., 1999), we make substantially more errors when presented with unfamiliar faces (Bruce et al., 1999, 2001; Henderson et al., 2001; Megreya & Burton, 2006, 2008). This is because the latter context relies on decision-making that is closely bound to the visual properties of specific images (Hancock et al., 2000), rather than the use of prior knowledge regarding the idiosyncratic variability inherent in each individual's facial appearance across images and situations (Burton et al., 2016). Given that facial information is necessarily limited when presented with one or a few images of unfamiliar identities, researchers have found it difficult to establish methods of improving performance on unfamiliar face matching tasks.

One route to improvement may be to consider the nature of the stimuli presented. Rather than a single, passport-style facial photograph, studies have suggested that providing either an average image derived from several original photographs (White et al., 2014) or those multiple images themselves (Menon et al., 2015; White et al., 2014) can increase face matching performance. However, more recent work has found no benefits from the use of these approaches (Kramer &

Reynolds, 2018; Ritchie et al., 2018, 2020, 2021; Sandford & Ritchie, 2021).

A second route to improvement has targeted the individual completing the task. Researchers have considered whether specific training or instruction may result in increased accuracy. However, professional training programs that are currently in use fail to improve performance (Towler et al., 2019). In the laboratory, instructing participants to utilise a feature-by-feature comparison strategy has garnered mixed results (Megreya, 2018; Megreya & Bindemann, 2018), perhaps due to the lack of clarity regarding which features should be prioritised. More recently, Towler and colleagues (Towler et al., 2021; see also Carragher et al., 2022) showed that a training intervention where participants were instructed to prioritise the ears and facial marks (most useful according to professional examiners; Towler et al., 2017) led to performance improvements. However, these improvements were absent in a follow-up study (Kramer, 2023).

Rather than instructing or training the individual regarding their approach to the task, a simple way to produce higher accuracy appears to result from working in pairs. When two people completed an unfamiliar face matching task together, their performance was better than when each completed the task individually (Dowsett & Burton, 2015; Ritchie et al., 2022). Interestingly, the 'social' aspect of this pairing is unnecessary in that non-social pairs (i.e., individuals who completed the task alone, and are later paired 'synthetically' by averaging responses across two participants) also outperformed individuals (Balsdon et al., 2018; White et al., 2013, 2015) and, indeed, showed performance levels similar to social pairs (Cavazos et al., 2023; Jeckeln et al., 2018). The process of averaging the responses of two or more individuals has been termed 'wisdom of the crowd', a phenomenon first noted by Galton (1907), and makes use of the fact that aggregating across individuals minimises the influences of both random noise and idiosyncratic biases.

More recently, researchers have begun to investigate the possible benefits associated with the 'wisdom of the inner crowd' (see Herzog & Hertwig, 2014). Rather than averaging the responses of two (or more) individuals (below, termed the outer crowd), this inner crowd refers to the aggregation of multiple responses given by the same person. Although this technique is unable to remove idiosyncratic biases, aggregated responses should still be more accurate than individual responses due to the decrease in random noise. Indeed, several studies have now demonstrated this result across different tasks (e.g., Hourihan & Benjamin, 2010; Steegen et al., 2014; van Dolder & van den Assem, 2018; Vul & Pashler, 2008), although this has yet to be investigated within the domain of face matching.

Evidence suggests that an inner crowd of two (i.e., the same person provides two answers to the same question) performs poorer than an outer crowd of the same size (that is, aggregating the answers given by two different people) for the reasons noted above (van Dolder & van den Assem, 2018; Vul & Pashler, 2008). However, inner crowds (within-person aggregation) may still be useful in specific, applied contexts in which only one individual is available to provide responses. For example, in the context of face identification, it may be more cost effective or practical to collect multiple responses from a single forensic examiner rather than obtaining responses from two examiners. As such, when outer crowds are unavailable as a solution, the use of an inner crowd could be beneficial.

Although we have introduced inner crowds as a promising route to the improvement of face matching performance, the same reasoning can also be applied to voice matching. Within a forensic context, unfamiliar voice matching may play a role in situations where perpetrators are encountered under poor visual conditions or when an offence is committed over the telephone. Researchers have identified large individual differences in voice matching and identification abilities (Lavan et al., 2019; Mühl et al., 2018; Smith et al., 2019, 2020; Sunilkumar et al., 2023), in addition to creating several standardised tests of unfamiliar voice matching (Mühl et al., 2018; Sunilkumar et al., 2023). However, to date, we are unaware of any research attempting to improve performance on such tasks.

1.1 | The current study

Previous research on unfamiliar face matching found that outer crowds produced higher performance than individuals. However, in some situations, there may only be one individual available with the expertise to provide judgments, with inner crowds representing the only option. Here, we considered whether inner crowds also performed better than participants' single test responses.

While the current study involved participants completing the same test twice, with no interval in between tests, we also reanalysed previously collected data (Kramer et al., 2021) resulting from the same experimental design but with at least a one-week interval between testing sessions. As such, we were able to investigate whether any benefits associated with inner crowds required an interval between responses.

Finally, along with face matching, we also addressed these research questions using an unfamiliar voice matching test. To date, there has been no consideration of either outer or inner crowd benefits within this domain, and so we explored whether voice matching could be improved through one or both of these routes.

2 | EXPERIMENT 1

In this first experiment, unfamiliar face matching was assessed twice, with no interval between assessments, using the same 40-trial test. We compared performance in each test with responses derived from both inner and outer crowds. We then reanalysed previously collected data (Kramer et al., 2021), which followed the same procedure, using the same test, but incorporated at least a one-week gap between assessments.

2.1 | Method

2.1.1 | Participants

A sample of 52 volunteers (38 women, 14 men; age $M = 32.9$ years, $SD = 16.2$ years; 92% self-reported ethnicity as White) provided written, informed consent online before taking part, and received an onscreen debriefing upon completion of the experiment.

The sample size was based on our plan to compare participants' two original performances on the test (referred to below as T1 and T2) with responses calculated from their inner crowds. As such, the use of a one-way repeated measures analysis of variance (ANOVA) with three measurements meant that a sample size of at least 28 was required to detect a medium-sized effect after choosing an α of .05 and with power set to 0.80 (G*Power 3.1 software; Faul et al., 2007).

In addition, to compare inner and outer crowd performance using a between-participants *t*-test, and with the aim of detecting a medium-sized effect after choosing an α of .05 and with power set to 0.80, we required a sample size of at least 34 participants. With an allocation ratio (N2/N1) of 16, this meant that the comparison group size needed to be at least 536, which is lower than the 561 outer crowds that would be produced (i.e., the number of possible simulated pairs of participants when choosing two people from 34). For sample sizes larger than 34, the number of ways of choosing two people from the sample would cause the allocation ratio to increase further, resulting in larger samples achieving power of at least 0.80.

Both experiments reported here were approved by the University of Lincoln's ethics committee (ref. 13627) and were carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki.

2.1.2 | Stimuli

We used the short version of the Glasgow face matching test (GFMT; Burton et al., 2010) to assess performance. The task comprised 40 pairs of adult male (24) and female faces (16), where half of the pairs were match trials (different images of the same person) and half were mismatch trials (different people with a similar appearance). All images were greyscale, passport-style photographs, depicting a front-on, neutral expression, and displayed on a plain, white background.

2.1.3 | Procedure

The experiment was completed online using the Qualtrics survey platform (www.qualtrics.com). After consent was obtained, participants provided demographic information (age, gender, and ethnicity). Through the information provided onscreen at the start of the experiment, participants were informed that they would be completing the same task twice.

First, participants completed all 40 trials of the short version of the GFMT (referred to as T1). On each trial, two face photographs were displayed onscreen and participants were instructed to decide whether they thought these faces were the same person or two different people. Following previous research (Kramer et al., 2021; O'Toole et al., 2007), responses were provided using a labelled rating scale: (1) sure they are the same; (2) think they are the same; (3) don't know; (4) think they are not the same; (5) sure they are not the same. Trial order was randomised for each participant, no time limits were imposed upon responses, and no feedback was given at any stage.

Upon completion of the test, participants were immediately presented with an instruction screen explaining that they would next complete the same test for a second time. This second presentation of the test (referred to as T2) was identical to the first (i.e., the same 40 face matching trials), with trial order again randomised for each participant. As before, no time limits were imposed upon responses and no feedback was given.

2.2 | Results

We first analysed the current set of data, where the GFMT was completed twice with no interval between tests. Following this, we reanalysed the data from Kramer et al.'s (2021) Experiment 1, where the GFMT was also completed twice by participants, but with tests separated by at least 1 week.

2.2.1 | Current experiment—no interval between tests

Before considering questions regarding inner and outer crowds, we first calculated some properties of the GFMT data. The internal reliabilities at T1 (Cronbach's $\alpha = .98$) and T2 (Cronbach's $\alpha = .99$) were high, as was the test–retest reliability of the AUC values, $r = .71$. In addition, collapsing across all participants and trials, we summarised the confidence ratings given at T1 ($M = 3.00$, $SD = 1.51$) and T2 ($M = 2.77$, $SD = 1.54$).

Rather than making explicit, binary judgements about whether face pairs were the 'same' or 'different' people, participants rated the likelihood that the two images were of the same person (e.g., Kramer et al., 2021; O'Toole et al., 2007). This approach meant that representational and decisional components could be separated, with the focus being placed on the former. The use of response scales in applied settings allows for the decision threshold to be varied in response to varying risk associated with different decisions, for example, if it is desirable to avoid 'miss' decisions (match trials given a 'different' response) then the threshold for 'same' responses could be set lower than if the priority is to avoid 'false alarms'.

For each participant, separately for each completion of the test, we calculated the hit and false alarm rates for each possible threshold along the rating scale (1 through 5). Plotting these values produced the receiver operating characteristic (ROC), with the area under this ROC curve (AUC) representing a measure that is widely used to assess the performance of classification rules over the entire range of possible thresholds (Krzanowski & Hand, 2009). As such, AUC allowed us to quantify the performance of a classifier (here, our participants), irrespective of where the cut-off between binary 'same'/'different' responses might have been placed. Here, we used AUC to quantify the extent to which ratings discriminated between match and mismatch trials (e.g., White et al., 2013).

Next, for each participant, we calculated the AUC value that resulted from their inner crowd. Simply, given that each trial received a rating (from 1 to 5) for each completion of the test, we calculated the (arithmetic) mean rating for that trial across the two tests and then used these aggregated trial ratings (collapsing across all participants and trials: $M = 2.89$, $SD = 1.41$) to calculate AUCs (e.g., Jeckeln et al., 2018). This inner crowd performance was compared with the original AUCs produced by each individual test using a one-way repeated measures ANOVA. We found a significant main effect, $F(2, 102) = 11.09$, $p < .001$, $\eta_p^2 = 0.18$, with pairwise comparisons (Bonferroni corrected) revealing that inner crowds ($M = 0.92$, 95% CI [0.89, 0.95]) performed significantly better than individuals at both T1 ($M = 0.87$, 95% CI [0.85, 0.90]; $p < .001$) and T2 ($M = 0.89$, 95% CI [0.86, 0.92]; $p = .002$). Performance at T1 and T2 did not differ ($p = .658$; see Figure 1).

Finally, we simulated the performance of participant pairs (also called 'non-social dyads'; for example, Balsdon et al., 2018; Davis et al., 2019; Jeckeln et al., 2018; White et al., 2013), an example of an outer crowd of two people. For every possible pairing of participants (i.e., the 1326 ways of choosing two people from the sample), we calculated the mean rating for each trial (using individuals' T1 responses) and the resulting AUC. An independent-samples *t*-test (with Levene's test, $p < .001$, meaning that equal variances were not assumed) found that outer crowds ($M = 0.94$, 95% CI [0.94, 0.94]) showed no difference in performance in comparison with inner crowds: $t(52) = 1.57$, $p = .122$, Cohen's $d = .22$ (see Figure 1).

2.2.2 | Reanalysis of Kramer et al. (2021)—One week interval between tests

Although Kramer et al.'s (2021) experiments focussed on the consistency in responses given by the same participants to the same trials across two different testing sessions, the nature of the dataset meant that we could also investigate whether wisdom of the inner crowd might result in improvements to performance.

Fifty participants completed the short version of the GFMT (Burton et al., 2010; also used in the current experiment) on two separate occasions, with an interval of at least 1 week (range = 12–35 days) between testing sessions. Participants responded using a 1–5 scale, identical to the one used in the current experiment. The internal reliabilities at T1 (Cronbach's $\alpha = .99$) and T2 (Cronbach's $\alpha = .99$) were high, as was the test–retest reliability of the AUC values, $r = .58$. In addition, collapsing across all participants and trials, we summarised the confidence ratings given at T1 ($M = 2.89$, $SD = 1.59$) and T2 ($M = 2.85$, $SD = 1.58$), as well as for the inner crowds ($M = 2.87$, $SD = 1.45$).

First, we calculated inner crowd performance for each participant, and compared these values to individual test performance using a one-way repeated measures ANOVA. We found a significant main effect, $F(2, 98) = 13.29$, $p < .001$, $\eta_p^2 = 0.21$, with pairwise comparisons (Bonferroni corrected) revealing that inner crowds ($M = 0.94$, 95% CI [0.93, 0.96]) performed significantly better than individuals at both T1 ($M = 0.90$, 95% CI [0.87, 0.92]; $p < .001$) and T2 ($M = 0.91$, 95% CI [0.88, 0.94]; $p < .001$). Performance at T1 and T2 did not differ ($p = 1.00$; see Figure 2).

Next, we simulated the performance of participant pairs for every possible pairing (i.e., the 1225 ways of choosing two people from the sample) and compared these outer crowd AUCs with inner crowd performance. An independent-samples *t*-test (with Levene's test, $p < .001$, meaning that equal variances were not assumed) found no difference between outer ($M = 0.96$, 95% CI [0.95, 0.96]) and inner crowds: $t(51) = 1.49$, $p = .142$, Cohen's $d = 0.21$ (see Figure 2).

2.3 | Discussion

The results of this experiment demonstrated that performance was significantly increased by using the wisdom of the inner crowd. Simply averaging two responses, given by the same individual, led to a 5% improvement in AUC values in comparison with participants' original (T1) responses. In line with previous work (e.g., White et al., 2013,

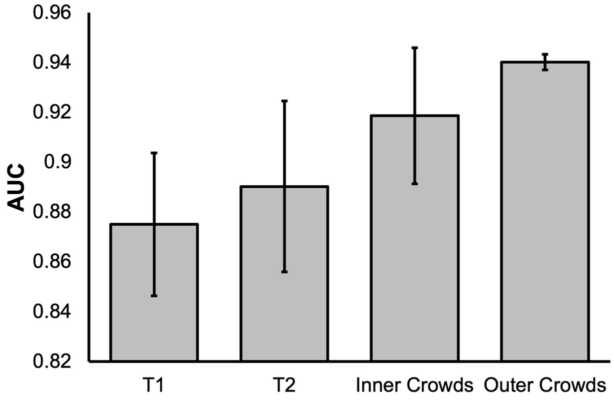

FIGURE 1 Performance on the GFMT with no interval between the two tests (T1 and T2). Error bars represent 95% confidence intervals.

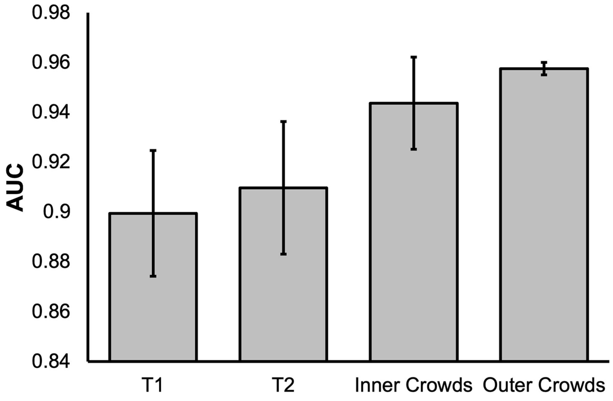

FIGURE 2 Performance on the GFMT with a one-week interval between the two tests (T1 and T2). Error bars represent 95% confidence intervals.

2015), we found that outer crowds of two people also produced performance benefits (of around 6%–7%) over individuals. Interestingly, these findings were evident when no interval, as well as a one-week interval, appeared between testing sessions.

3 | EXPERIMENT 2

While Experiment 1 demonstrated that both inner and outer crowds produced significant increases in face matching performance, there has yet to be any consideration of these approaches with voice matching. In this second experiment, unfamiliar voice matching was assessed twice, with no interval between assessments, using the same 80-trial test. We compared performance in each test with responses derived from both inner and outer crowds. We then reanalysed previously collected data (Kramer et al., 2021), which followed the same procedure, using the same test, but incorporated at least a one-week gap between assessments.

3.1 | Method

3.1.1 | Participants

A new sample of 47 volunteers (31 women, 15 men, 1 nonbinary; age $M = 42.9$ years, $SD = 14.9$ years; 96% self-reported ethnicity as White) provided written, informed consent online before taking part, and received an onscreen debriefing upon completion of the experiment. There was no overlap between this sample and those who participated in Experiment 1.

As in Experiment 1, we required a sample size of at least 34 participants to detect medium-sized effects in both our one-way repeated measures ANOVA and between-participants *t*-test.

3.1.2 | Stimuli

We used the Bangor voice matching test (BVMT; Mühl et al., 2018) to assess performance. The task comprised 80 pairs of adult male (40) and female voices (40), where half of the pairs were match trials (different voice samples produced by the same person) and half were mismatch trials (voice samples produced by two different people). Each sample was either a consonant-vowel-consonant (e.g., 'had') or vowel-consonant-vowel (e.g., 'aba').

3.1.3 | Procedure

The experiment was completed online using the Gorilla experiment builder (Anwyl-Irvine et al., 2020). As in Experiment 1, we collected information regarding the participant's age, gender and ethnicity. Through the information provided onscreen at the start of the experiment, participants were informed that they would be completing the same task twice.

First, participants completed all 80 trials of the BVMT (T1). On each trial, two buttons were displayed onscreen (labelled 'Play Sound 1' and 'Play Sound 2') and participants were instructed to decide whether they thought these audio samples were produced by the same speaker or different speakers. As in Experiment 1, responses were provided using a 1–5 rating scale. Participants were able to listen to the voice samples an unlimited number of times, by clicking on the two buttons, prior to giving their response. Between trials, a fixation cross appeared for 800 ms. Trial order was randomised for each participant, no time limits were imposed upon responses, and no feedback was given at any stage.

Upon completion of the test, participants were immediately presented with an instruction screen explaining that they would next complete the same test for a second time. This second presentation of the test (T2) was identical to the first (i.e., the same 80 voice matching trials), with trial order again randomised for each participant. As before, no time limits were imposed upon responses and no feedback was given.

3.2 | Results

We first analysed the current set of data, where the BVMT was completed twice with no interval between tests. Following this, we reanalysed the data from Kramer et al.'s (2021) Experiment 2, where the BVMT was also completed twice by participants, but with tests separated by at least 1 week.

3.2.1 | Current experiment—no interval between tests

Before considering questions regarding inner and outer crowds, we first calculated some properties of the BVMT data. The internal reliabilities at T1 (Cronbach's $\alpha = .97$) and T2 (Cronbach's $\alpha = .97$) were high, as was the test–retest reliability of the AUC values, $r = .81$. In addition, collapsing across all participants and trials, we summarised the confidence ratings given at T1 ($M = 2.66$, $SD = 1.51$) and T2 ($M = 2.62$, $SD = 1.60$), as well as for the inner crowds ($M = 2.64$, $SD = 1.37$).

Next, for each participant, we calculated the AUC value that resulted from their inner crowd. Using a one-way repeated measures ANOVA, this inner crowd performance was compared with the original AUCs produced by each individual test. We found a significant main effect, $F(2, 92) = 26.74$, $p < .001$, $\eta_p^2 = 0.37$, with pairwise comparisons (Bonferroni corrected) revealing that inner crowds ($M = 0.85$, 95% CI [0.82, 0.88]) performed significantly better than individuals at both T1 ($M = 0.80$, 95% CI [0.76, 0.83]; $p < .001$) and T2 ($M = 0.81$, 95% CI [0.78, 0.84]; $p < .001$). Performance at T1 and T2 did not differ ($p = .511$; see Figure 3).

Finally, we simulated the performance of participant pairs for every possible pairing (i.e., the 1081 ways of choosing two people from the sample) and compared these outer crowd AUCs with inner

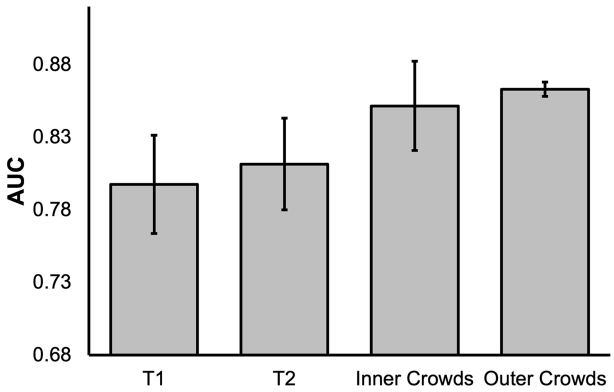

FIGURE 3 Performance on the BVMT with no interval between the two tests (T1 and T2). Error bars represent 95% confidence intervals.

crowd performance. An independent-samples *t*-test (with Levene's test nonsignificant, $p = .125$) found that outer crowds ($M = 0.86$, 95% CI [0.86, 0.87]) showed no difference in performance in comparison with inner crowds: $t(1126) = 0.93$, $p = .355$, Cohen's $d = 0.14$ (see Figure 3).

3.2.2 | Reanalysis of Kramer et al. (2021)—One week interval between tests

Forty-five participants completed the BVMT (Mühl et al., 2018; also used in the current experiment) on two separate occasions, with an interval of at least 1 week (range = 11–34 days) between testing sessions. Participants responded using a 1–5 scale, identical to the one used in the current experiment. The internal reliabilities at T1 (Cronbach's $\alpha = .97$) and T2 (Cronbach's $\alpha = .97$) were high, as was the test–retest reliability of the AUC values, $r = .58$. In addition, collapsing across all participants and trials, we summarised the confidence ratings given at T1 ($M = 2.70$, $SD = 1.56$) and T2 ($M = 2.59$, $SD = 1.53$), as well as for the inner crowds ($M = 2.64$, $SD = 1.34$).

First, we calculated inner crowd performance for each participant, and compared these values to individual test performance using a one-way repeated measures ANOVA. We found a significant main effect, $F(2, 88) = 14.05$, $p < .001$, $\eta_p^2 = 0.24$, with pairwise comparisons (Bonferroni corrected) revealing that inner crowds ($M = 0.87$, 95% CI [0.84, 0.90]) performed significantly better than individuals at both T1 ($M = 0.83$, 95% CI [0.81, 0.86]; $p < .001$) and T2 ($M = 0.81$, 95% CI [0.77, 0.85]; $p < .001$). Performance at T1 and T2 did not differ ($p = .329$; see Figure 4).

Next, we simulated the performance of participant pairs for every possible pairing (i.e., the 990 ways of choosing two people from the sample) and compared these outer crowd AUCs with inner crowd performance. An independent-samples *t*-test (with Levene's test, $p < .001$, meaning that equal variances were not assumed) found no difference between outer ($M = 0.90$, 95% CI [0.89, 0.90]) and inner crowds: $t(45) = 2.00$, $p = .052$, Cohen's $d = .30$ (see Figure 4).

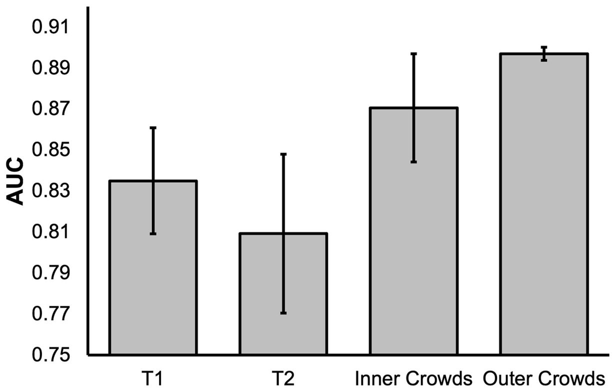

FIGURE 4 Performance on the BVMT with a one-week interval between the two tests (T1 and T2). Error bars represent 95% confidence intervals.

3.3 | Discussion

The results of this second experiment mirrored those of Experiment 1. Performance was significantly increased by using the wisdom of the inner crowd (of around 4%–7%), as well as the outer crowd (of around 7%–8%). Again, these findings were evident when no interval, as well as a one-week interval, appeared between testing sessions.

4 | GENERAL DISCUSSION

Through focussing on interventions regarding either the stimuli or the individual, previous research has failed to identify robust methods of performance improvement within the domain of face matching (e.g., Kramer, 2023; Kramer & Reynolds, 2018; Ritchie et al., 2018). However, by aggregating responses across individuals to produce outer crowds, significant accuracy benefits have already been shown (e.g., Cavazos et al., 2023; Jeckeln et al., 2018). Here, we considered again the possibility of this outer crowd benefit for face matching, while also investigating performance on a voice matching task. In addition, for both domains, we explored whether inner crowds (i.e., aggregating responses given by the same individual) might produce similar performance improvements.

Both experiments presented here revealed the same pattern of results. First, inner crowds performed better than individual responses given in either test (T1 or T2). Second, outer crowds were also beneficial but did not outperform inner crowds. Third, these results were evident when no gap was included between tests, as well as when a one-week interval was incorporated between testing sessions.

While inner crowds have not been investigated previously with regard to face or voice matching, performance increases have been shown in other domains (e.g., real-world knowledge—Vul & Pashler, 2008; object number estimation—van Dolder & van den Assem, 2018). However, in those tasks, inner crowds provided far less of a benefit than outer crowds. In fact, averaging around 10 responses from the same individual was shown to provide the same benefit as averaging estimates from two different people (van Dolder & van den Assem, 2018; Vul & Pashler, 2008). In contrast, when considering the evaluation of college entrance essays, Barneron et al. (2019) found that, while outer crowds were more accurate than inner crowds (incorporating two responses in both cases), the difference in performance of the two methods of aggregation was fairly small. Here, we found no difference between our inner and outer crowds. It is worth noting that we were not sufficiently powered to detect small effect sizes with our comparisons and so it may be that outer crowds (of two people) could provide significant benefits over inner crowds, and this remains an avenue for future research.

Previous research has shown that inner crowd benefits tended to be greater with a delay between estimates (Steegen et al., 2014; van Dolder & van den Assem, 2018; Vul & Pashler, 2008). By analysing data from the current experiments, along with a reanalysis of previously collected data (Kramer et al., 2021), we found that inner crowds performed better than individual responses both without a delay and with a one-week delay. While the sample sizes involved here did not allow for a direct comparison between these conditions (which would require 64 participants in each sample to detect a medium-sized effect with 80% power), it may be that the length of the matching tests used (40 face trials and 80 voice trials) prevented a delay from providing additional benefits. Over so many trials, participants would likely have little recollection of their first responses when it came to completing the test for a second time. In comparison, previous research demonstrating the benefits associated with a delay between responses utilised far shorter tasks: eight questions probing real-world knowledge (Steegen et al., 2014; Vul & Pashler, 2008) and a single guess as to the number of objects in a transparent container (van Dolder & van den Assem, 2018). In these situations, asking individuals to complete the task twice without a delay would likely see respondents either reproduce their initial answers or be biased by them. However, the idea that test length played a role here remains to be investigated and represents an interesting path for further study.

In addition to the potential improvement to inner crowds resulting from a delay between responses, researchers have also considered the effect of eliciting responses in different ways. For instance, rather than simply asking individuals to provide a second estimate, inner crowds were improved when this second estimate was made from a disagreeing perspective (e.g., asking how 'a friend whose views and opinions are very different from yours' would answer; Van de Calseyde & Efendić, 2022), was provided as an estimate of public opinion (Fujisaki et al., 2023), or was a dialectical estimate (e.g., using a consider-the-opposite strategy; Herzog & Hertwig, 2009). In all cases, the aim has been to improve the diversity and independence of the two estimates. In the current experiments, participants were not instructed to generate their responses in a particular way, and so we might predict that inner crowd performance could be improved by utilising one of these strategies.

At present, few methods have resulted in robust improvements in face matching performance. While the training intervention designed by Towler et al. (2021) showed some promise, this approach subsequently failed to replicate (Kramer, 2023). Even so, their training produced an accuracy increase of only 6% in AUC post-intervention. Interestingly, our inner crowds showed around the same level of increase without the need for any type of training or intervention. Instead, by asking (untrained) individuals to provide a second response on each trial, an aggregate of these responses led to a comparable performance increase. Further, to our knowledge, there are currently no established routes to increased performance on tests of voice matching, and so our demonstration that both inner and outer crowds achieved significant increases represents an initial step in this direction.

It is worth highlighting that participants in both experiments showed no performance improvements when we compared their T1 and T2 results. That is, we found no practice effects in our data. While previous studies have shown a lack of improvement across repetitions of the same test using intervals of a day (Bindemann et al., 2012) or week (Kramer et al., 2021) between sessions, this was also the case in the current work, where no interval was included. Participants did not receive any feedback during or after each test, and so there is perhaps no reason to predict an improvement across sessions. Previous research has shown a high level of consistency in responses when repeating the same test twice (Kramer et al., 2021). This could be explained by participants either remembering and reproducing their responses, or simply giving the same responses when faced with the same stimuli. Interestingly, for face recognition tests (where memory plays a role), practice effects have been identified (Murray & Bate, 2020).

While the current set of experiments found no benefit for two-person outer crowds over inner crowds of two responses, previous research has established that, as outer crowds grow larger in size, they are expected to outperform inner crowds (Balsdon et al., 2018; White et al., 2013, 2015). However, it is important to consider situations in which the use of an outer crowd may and may not be a viable option. For example, in certain contexts, requiring a border force officer or forensic examiner to provide a second response (perhaps after a delay) may be more feasible logistically than asking others to provide their opinions for aggregation.

Related, when considering the practical applications of aggregating responses, we note that our use of a 1–5 rating scale mirrors forensic examiner responses. Of course, many real-world contexts require a clear decision regarding whether two stimuli originated from the same person (e.g., when identifying a traveller at border control) and so do not allow for a 'don't know' response. While evidence suggests that binary decisions can also benefit from the use of outer crowd aggregation (e.g., by considering the proportion of 'same' responses and applying a majority vote decision rule; White et al., 2013), further research might seek to confirm that inner crowds would result in improved performance when limiting responses to binary decisions, or perhaps a rating scale with an even number of options and no middle value.

In summary, for tests of both face and voice matching, we considered whether averaging two responses given by the same individual (wisdom of the inner crowd) would produce an increase in performance. Our results demonstrated that inner crowds significantly increased performance over participants' individual responses across both domains. Indeed, this gain in accuracy was no different than the well-known wisdom of the (outer) crowd (for crowds of two people). Finally, we found inner crowd benefits when tests were completed twice without a delay, as well as when a one-week interval was incorporated between testing sessions. We therefore propose that inner crowds may provide a simple and robust method of improving matching accuracy, in particular for forensic and security contexts in which outer crowds are either unavailable or impractical solutions.

ACKNOWLEDGMENTS

The authors thank Tia Beasley and Georgia Rutherford-Browne for help with participant recruitment.

CONFLICT OF INTEREST STATEMENT

There are no conflicts of interest to be disclosed.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Open Science Framework at https://osf.io/ezt7d/?view_only=3bd6de4f7a3a4f1892355003e782582d.

ORCID

Robin S. S. Kramer https://orcid.org/0000-0001-8339-8832

REFERENCES

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388–407.

Balsdon, T., Summersby, S., Kemp, R. I., & White, D. (2018). Improving face identification with specialist teams. *Cognitive Research: Principles and Implications*, *3*, 25.

Barneron, M., Allalouf, A., & Yaniv, I. (2019). Rate it again: Using the wisdom of many to improve performance evaluations. *Journal of Behavioral Decision Making*, *32*(4), 485–492.

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology: Applied*, *18*(3), 277–291.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*, 339–360.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*, 207–218.

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, *40*(1), 202–223.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, *42*(1), 286–291.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, *10*, 243–248.

Carragher, D. J., Towler, A., Mileva, V. R., White, D., & Hancock, P. J. (2022). Masked face identification is improved by diagnostic feature training. *Cognitive Research: Principles and Implications*, *7*(1), 30.

Cavazos, J. G., Jeckeln, G., & O'Toole, A. J. (2023). Collaboration to improve cross-race face identification: Wisdom of the multi-racial crowd? *British Journal of Psychology*. Advance online publication.

Davis, J. P., Maigut, A., & Forrest, C. (2019). The wisdom of the crowd: A case of post- to ante-mortem face matching by police super-recognisers. *Forensic Science International*, *302*, 109910.

Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology*, *106*(3), 433–445.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Fujisaki, I., Yang, K., & Ueda, K. (2023). On an effective and efficient method for exploiting the wisdom of the inner crowd. *Scientific Reports*, *13*, 3608.

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451.

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, *4*, 330–337.

Henderson, Z., Bruce, V., & Burton, A. M. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology*, *15*, 445–464.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*(2), 231–237.

Herzog, S. M., & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*, *18*(10), 504–506.

Hourihan, K. L., & Benjamin, A. S. (2010). Smaller is better (when sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(4), 1068–1074.

Jeckeln, G., Hahn, C. A., Noyes, E., Cavazos, J. G., & O'Toole, A. J. (2018). Wisdom of the social versus non-social crowd in face identification. *British Journal of Psychology*, *109*(4), 724–735.

Kramer, R. S. S. (2023). Face matching and metacognition: Investigating individual differences and a training intervention. *PeerJ*, *11*, e14821.

Kramer, R. S. S., Jones, A. L., & Gous, G. (2021). Individual differences in face and voice matching abilities: The relationship between accuracy and consistency. *Applied Cognitive Psychology*, *35*(1), 192–202.

Kramer, R. S. S., & Reynolds, M. G. (2018). Unfamiliar face matching with frontal and profile views. *Perception*, *47*(4), 414–431.

Krzanowski, W. J., & Hand, D. J. (2009). *ROC curves for continuous data*. Chapman & Hall.

Lavan, N., Burston, L. F. K., & Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology*, *110*(3), 576–593.

Megreya, A. M. (2018). Feature-by-feature comparison and holistic processing in unfamiliar face matching. *PeerJ*, *6*, e4437.

Megreya, A. M., & Bindemann, M. (2018). Feature instructions improve face-matching accuracy. *PLoS One*, *13*(3), e0193455.

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*, 865–876.

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied*, *14*, 364–372.

Menon, N., White, D., & Kemp, R. I. (2015). Variation in photos of the same face drives improvements in identity verification. *Perception*, *44*, 1332–1341.

Mühl, C., Sheil, O., Jarutytė, L., & Bestelmeyer, P. E. (2018). The Bangor voice matching test: A standardized test for the assessment of voice perception ability. *Behavior Research Methods*, *50*(6), 2184–2192.

Murray, E., & Bate, S. (2020). Diagnosing developmental prosopagnosia: Repeat assessment using the Cambridge face memory test. *Royal Society Open Science*, *7*(9), 200884.

O'Toole, A. J., Phillips, P. J., Jiang, F., Ayyad, J., Penard, N., & Abdi, H. (2007). Face recognition algorithms surpass humans matching faces over changes in illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(9), 1642–1646.

Ritchie, K. L., Flack, T. R., Fuller, E. A., Cartledge, C., & Kramer, R. S. S. (2022). The pairs training effect in unfamiliar face matching. *Perception*, *51*(7), 477–495.

Ritchie, K. L., Kramer, R. S. S., Mileva, M., Sandford, A., & Burton, A. M. (2021). Multiple-image arrays in face matching tasks with and without memory. *Cognition*, *211*, 104632.

Ritchie, K. L., Mireku, M. O., & Kramer, R. S. S. (2020). Face averages and multiple images in a live matching task. *British Journal of Psychology*, *111*(1), 92–102.

Ritchie, K. L., White, D., Kramer, R. S. S., Noyes, E., Jenkins, R., & Burton, A. M. (2018). Enhancing CCTV: Averages improve face identification from poor quality images. *Applied Cognitive Psychology*, *32*, 671–680.

Sandford, A., & Ritchie, K. L. (2021). Unfamiliar face matching, within-person variability, and multiple-image arrays. *Visual Cognition*, *29*(3), 143–157.

Smith, H. M., Baguley, T. S., Robson, J., Dunn, A. K., & Stacey, P. C. (2019). Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance. *Applied Cognitive Psychology*, *33*(2), 272–287.

Smith, H. M., Bird, K., Roeser, J., Robson, J., Braber, N., Wright, D., & Stacey, P. C. (2020). Voice parade procedures: Optimising witness performance. *Memory*, *28*(1), 2–17.

Steegen, S., Dewitte, L., Tuerlinckx, F., & Vanpaemel, W. (2014). Measuring the crowd within again: A pre-registered replication study. *Frontiers in Psychology*, *5*, 786.

Sunilkumar, D., Kelly, S. W., Stevenage, S. V., Rankine, D., & Robertson, D. J. (2023). Sounds and speech: Individual differences in unfamiliar voice recognition. *Applied Cognitive Psychology*, *37*, 507–519.

Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., & White, D. (2019). Do professional facial image comparison training courses work? *PLoS One*, *14*(2), e0211037.

Towler, A., Keshwa, M., Ton, B., Kemp, R. I., & White, D. (2021). Diagnostic feature training improves face matching accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(8), 1288–1298.

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied*, *23*(1), 47–58.

Van de Calseyde, P. P. F. M., & Efendić, E. (2022). Taking a disagreeing perspective improves the accuracy of people's quantitative estimates. *Psychological Science*, *33*(6), 971–983.

van Dolder, D., & van den Assem, M. J. (2018). The wisdom of the inner crowd in three large natural experiments. *Nature Human Behaviour*, *2*(1), 21–26.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647.

White, D., Burton, A. M., Jenkins, R., & Kemp, R. (2014). Redesigning photo-ID to improve unfamiliar face matching performance. *Journal of Experimental Psychology: Applied*, *20*, 166–173.

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. *Applied Cognitive Psychology*, *27*(6), 769–777.

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences*, *282*, 20151292.

How to cite this article: Kramer, R. S. S., Flanagan, E., Jones, A. L., & Gous, G. (2023). Wisdom of the inner crowd benefits both face and voice matching. *Applied Cognitive Psychology*, 1–9. https://doi.org/10.1002/acp.4133