**ACCEPTED VERSION OF MANUSCRIPT**

# Examples improve facial identity comparison

Natalie W. Gentry & Markus Bindemann

School of Psychology, University of Kent, UK

Correspondence to:

Natalie Gentry, School of Psychology, University of Kent, Canterbury CT2 7NP, UK.

Email: nataliewgentry@outlook.com

Word count of Introduction and General Discussion: 2404

Total word count (excluding abstract, references and figure captions): 4668

**Abstract**

Unfamiliar-face matching requires an identity comparison of two simultaneously presented faces that are unknown to the viewer. This can be a difficult task, even for police and security professionals who perform such comparisons routinely. This study examined whether the provision of example face pairs, presented either side of a target face pair and clearly labelled as identity-matches and mismatches, improves matching accuracy. Examples aided performance at group level, but analysis of individual differences indicates that this arises from improvement in lower-performing observers, who were initially least accurate. This effect generalised to previously unseen faces from the stimulus set from which target and example pairs were drawn, but not to face pairs from a new stimulus set. We suggest that examples aid performance by providing criteria to distinguish identity-matches from mismatches, which observers would otherwise have to deduce by their own judgement during participation.

Keywords: face matching, identity comparison, examples, improvement, individual differences, training

## Introduction

Unfamiliar-face matching requires a comparison of two side-by-side faces unknown to a viewer, to decide whether these depict the same person (an identity-match) or different people (a mismatch). This task is performed routinely in security settings, such as passport control, to monitor the movement of people across international borders. Despite its ubiquity in security settings, a coherent body of psychological research demonstrates that unfamiliar-face matching is prone to error (for reviews, see Fysh & Bindemann, 2017a; Robertson, Middleton, & Burton, 2015). Under idealised conditions, in which observers compare same-day high-quality photographs of faces, errors are made on 10-20% of trials (Bindemann, Avetisyan, & Blackwell, 2010; Burton, White, & McNeill, 2010; Megreya & Burton, 2006). These error rates are considered problematic for large-scale security operations, where a small percentage of errors can result in a large number of cases that give rise to incorrect decisions (Dhir, Singh, Kumar, & Singh, 2010; Jenkins & Burton, 2008). Accuracy declines further under conditions likely to present in applied settings, such as when to-be-compared face photographs were taken many months apart, as is typically the case with a passport photograph and its bearer (Megreya, Sandford, & Burton, 2013), when faces are matched over extended time periods (Alenezi & Bindemann, 2013; Alenezi, Bindemann, Fysh, & Johnston, 2015), when identity mismatches occur infrequently (Papesh & Goldinger, 2014), when operatives are under time pressure to perform this task (Bindemann, Fysh, Cross, & Watts, 2016; Fysh & Bindemann, 2017b; Özbek & Bindemann, 2011; Wirth & Carbon, 2017), and during human supervision of automated facial recognition decisions (Fysh & Bindemann, 2018a; White, Dunn, Schmid, & Kemp, 2015).

These findings highlight that face matching is generally challenging, but are based on measures of mean performance, across groups of participants. However, substantial individual differences also exist in this task, whereby some people perform close to chance when others

achieve high accuracy (Bindemann, Avetisyan, & Rakow, 2012; Burton et al., 2010; Fysh & Bindemann, 2018b). This performance range is important for demonstrating that security could be enhanced by selecting individuals with a specific aptitude for face processing (Bindemann et al., 2012; Bobak, Dowsett, & Bate, 2016; Bobak, Hancock, & Bate, 2016; White, Kemp, Jenkins, Matheson, & Burton, 2014). These individual differences indicate also that many face-matching errors do not arise from data limits, whereby stimuli carry insufficient information for accurate identifications to be made, but from the failure of some observers to correctly use the available information (Fysh & Bindemann, 2017a; Jenkins & Burton, 2011). In turn, the observation that other people can successfully match the same stimuli indicates that improvements in accuracy are in principle possible for those people who do not perform to a high level.

Limited research still exists on methods to improve a person's face-matching accuracy, and not all methods procure benefits. Training observers to classify face shapes, for example, does not improve face-matching accuracy (Towler, White, & Kemp, 2014). Similarly, training courses provided for professionals in real-world security settings appeared to be limited in effectiveness (Towler, Kemp, Burton, Dunn, Wayne, Moreton, & White, 2019). One method that appears to improve face matching, however, is feedback for whether a correct decision has been made. Under experimental conditions, provision of feedback immediately after a face-matching trial can help to maintain accuracy in subsequent trials of this task (Alenezi & Bindemann, 2013), and feedback can *improve* subsequent performance when this is provided whilst a just-classified face pair is still in view (White, Kemp, Jenkins, & Burton, 2014). However, real-world scenarios do not allow provision of feedback in this way, as the accuracy of matching decisions

is typically not known at the point of an identification. Consequently, observers rarely have the opportunity to learn from their face-matching decisions (Jenkins & Burton, 2011).

In this study, we investigate an alternative form of 'feedback' that could be provided in applied settings, to determine if this confers improvements in face-matching accuracy. Our approach is based on providing labelled example face-pairs of identity matches and mismatches, to the left and right of a centrally-presented target face-pair. The rationale for this manipulation is that matching errors may arise because observers do not have clearly defined *criteria* for distinguishing same- and different-identity face pairs. The observation that trial-by-trial feedback improves accuracy supports this reasoning and suggests that the feedback benefit arises by helping observers to refine their face-matching criteria.

In contrast to the trial-by-trial feedback manipulations of previous studies (Alenezi & Bindemann, 2013; White, Kemp, Jenkins, & Burton, 2014), the examples manipulation investigated here has greater potential to be implemented in applied settings because it does not require prior knowledge of the nature of the target face pair. To determine if such a benefit is found, we first assessed observers' face-matching accuracy without examples, to obtain a baseline measure of performance. We then provided examples in a second block of trials to improve performance. We also compared observers in this condition with another group, who were not provided with examples in the second block, on a between-subjects basis.

This group-level comparison provides a useful contrast to assess the *general* impact of examples on face-matching performance but, in line with other methods that have been investigated to improve accuracy in this task, we expected any gains at this level to be small (e.g., Alenezi & Bindemann, 2013; Megreya & Bindemann, 2018; Towler et al., 2014; White et al., 2014). However, considering the broad differences that exist in face-matching accuracy

between observers (Bindemann et al., 2012; Bobak, Pampoulov, & Bate, 2016; Burton et al., 2010; White, Kemp, Jenkins, Matheson, et al., 2014; for a review, see Lander, Bruce, & Bindemann, 2018), we were particularly interested in how examples influenced individual accuracy, by comparing any changes in performance with a person's baseline performance. Previous research shows, for example, that improvement with performance feedback is driven specifically by observers who initially display low aptitude in face matching (White et al., 2014). If the same applies to the provision of examples here, then one might also expect observers with low matching ability to improve the most.

If a matching improvement with the provision of examples is found, then it is also important to determine whether this transfers to conditions in which examples are no longer seen. To address this question, a subgroup of participants completed two additional blocks after the examples were withdrawn. One of these blocks comprised a repetition of the target face pairs from the examples block, but these were now shown without the example stimuli. The other block presented new target face pairs, which had not been encountered before in the experiment, but were taken from the same stimulus set. In addition, we also sought to assess whether any examples advantage would generalise to a completely different set of face stimuli. For this purpose, a second subgroup was also given a block of trials from a different face-matching test.

## Method

### *Participants*

One hundred and eighty students (140 female), with a mean age of 21.5 years (*SD* = 5.8; range: 18-57), took part in this experiment for course credit. All participants were of Caucasian ethnicity and reported normal or corrected-to-normal vision.

*Stimuli*

One hundred and twenty face pairs from the Glasgow Face Matching Test (GFMT) were employed as stimuli in this study (see Burton et al., 2010). These comprised of 60 identity-matches, in which two different same-day photographs of the same person were shown, and 60 identity-mismatches, depicting two different individuals in each pair. All the faces were depicted in greyscale, a frontal pose, and with a neutral expression, and were cropped to remove extraneous background. The maximum size for a face was 43 x 54 mm, while the maximum gap between faces in a pair was 25 mm. Each face pair was shown beneath the question "Match or Mismatch?".

In the experiment, 40 of these face pairs (20 matches, 20 mismatches) were employed as target stimuli, to assess observers' initial performance in Block 1, and were then repeated as the as target stimuli in Block 2 to measure improvement. Repeating the stimuli in this way ensured that any changes in *individual* performance across blocks cannot be attributed to variation in stimulus content. In addition, another 40 face pairs were presented as example stimuli to the left and right of the target pairs in Block 2 of the experimental condition. Two example face pairs were provided with each target stimulus, and were clearly labelled as identity-matches and mismatches. These example face pairs were randomly selected, but each pair occurred with equal frequency during the experiment. In addition, the sex of the example faces always matched that of the target face pair. For an illustration of a stimulus array, see Figure 1.

The remaining 40 face pairs were used to assess generalisation of improvement in a subgroup of participants, who were shown these stimuli without examples. In addition, a further block of 40 trials was included comprising 20 match and 20 mismatch stimuli from the Kent

Face Matching Test (KFMT; Fysh & Bindemann, 2018b). These face pairs consist of a relatively uncontrolled image from student photo-ID alongside a portrait that was recorded under controlled conditions. These different photographs were taken several months apart for each identity and vary in terms of, for example, hairstyle and expression (for full details, see Fysh & Bindemann, 2018b). In contrast to the face pairs from the GFMT, the KFMT stimuli therefore capture greater variation in appearance within identities. In this experiment, the KFMT portraits were presented at a size of 63 x 65 mm, whereas the photo-ID photographs measured 32 x 36 mm. Both photos were displayed on a blank white canvas, 65 mm apart.

*Procedure*

The experimental design is illustrated in Figure 2 and was implemented using 'PsychoPy' software (Peirce, 2007). All participants (N = 180) were shown two blocks, each containing the same 40 target face pairs (20 match and 20 mismatch) displayed in a randomly intermixed order. Half of the participants (N = 90) were assigned to the no-examples (control) condition while the remainder (N = 90) were assigned to the examples condition. In the no-examples condition, participants viewed only the target pairs in both of Block 1 and 2. In the examples condition, Block 1 was identical to the no-examples condition, but during the repetition of the target face pairs in Block 2, these were flanked by an example match and a mismatch face pair to the left and right. All participants were instructed to classify the centrally-presented target face pairs as identity-matches or mismatches as accurately as possible, by pressing one of two response keys on a standard computer keyboard. In addition, participants in the example condition were given additional instructions prior to Block 2, which explained the presence of the examples and encouraged observers to make use of these to aid their identification decisions. Specifically,

observers were informed that "on each screen you will be shown a pair of images of faces as before. This time you will be given an example of a match and a mismatch pair to help you make your decision." After each trial of Block 2, these participants were also asked to indicate whether they had made use of the examples to aid their last decision, by pressing one of two response keys on the computer keyboard.

In addition, a subgroup of participants from the examples and no-examples conditions (N = 60 each) was then given two further blocks. The first of these comprised a repetition of the stimuli from Block 1, to determine whether any increases in face-matching accuracy from viewing examples were retained when these were no longer present (here referred to as Block GFMT Old). In addition, the participants were either shown a block of 40 previously unseen face pairs (20 matches and 20 mismatches) from the GFMT (N = 30; referred to as Block GFMT New), or a block of 40 trials was included comprising 20 match and 20 mismatch stimuli from the KFMT (N = 30; referred to as Block KFMT). The order of these additional blocks (GFMT Old versus GFMT New / KFMT) was counterbalanced across participants, and trial order was randomised in all blocks.

## Results

*Group-level accuracy*

To determine the effect of examples on face matching at a group level, a 2 (condition: examples vs. no-examples) x 2 (block: Block 1 vs. Block 2) x 2 (trial type: match vs. mismatch) mixed-factor Analysis of Variance (ANOVA) was conducted. These data are illustrated in Figure 3 and reveal an interaction of block and condition, $F(1,178) = 6.51$, $p < .05$, $\eta_p^2 = 0.04$. Analysis of simple main effects shows that overall accuracy was similar for the examples condition and the no-examples condition in Block 1 (91.6 % and 92.5%), $F(1,178) = 0.63$, $p = .43$, $\eta_p^2 = 0.00$,

and Block 2 (93.9% and 92.6%), $F(1,178) = 1.64$, $p = .20$, $\eta_p^2 = 0.01$, and was also comparable across blocks in the no-examples condition (92.5% and 92.6%), $F(1,178) = 0.02$, $p = 0.89$, $\eta_p^2 = 0.00$. However, accuracy increased from Block 1 to Block 2 in the examples condition (91.6% and 93.9%), $F(1,178) = 14.03$, $p < .001$, $\eta_p^2 = 0.07$.

ANOVA also revealed a main effect of block, $F(1,178) = 7.56$, $p < .01$, $\eta_p^2 = 0.04$, and an interaction between block and trial type, $F(1,178) = 18.50$, $p < .001$, $\eta_p^2 = 0.09$. This interaction reflects that match and mismatch accuracy was comparable for Block 1 (91.1% and 93.0%), $F(1,178) = 2.96$, $p = .09$, $\eta_p^2 = 0.02$, and mismatch accuracy was also similar across both blocks (93.0% and 92.1%), $F(1,178) = 2.16$, $p = .14$, $\eta_p^2 = 0.01$. In contrast, match accuracy increased from Block 1 to Block 2 (91.1% and 94.3%), $F(1,178) = 22.65$, $p < .001$, $\eta_p^2 = 0.11$, and the difference in accuracy between match and mismatch trials was also approaching significance for Block 2 (94.3% and 92.1%), $F(1,178) = 3.86$, $p = .05$, $\eta_p^2 = 0.02$, due to higher accuracy on match trials.

Overall, these analyses therefore demonstrate that two separable effects occurred at group level. The first presents an improvement in face-matching accuracy with the provision of examples. The second suggests that the proportion of correct match responses increased across blocks over the course of the experiment. We note, however, that both effects were statistically reliable but numerically small, of 2.3% and 3.2%, respectively. None of the remaining main effects or interactions were significant, all $Fs \leq 2.06$, $ps \geq .15$, $\eta_p^2 \leq 0.01$.

*Individual differences*

To determine how examples affected performance at an individual level, observers' percentage accuracy in Block 1 was subtracted from Block 2 to provide a measure of change in

performance. This score was then correlated with Block 1 to determine whether any improvements in accuracy were related to individual differences in baseline performance. The correlations of baseline accuracy with improvement are summarized in Table 1, with the accuracy data illustrated in Figure 4. For the no-examples condition, a negative correlation was observed between baseline accuracy and the change in performance by Block 2 for match trials. This indicates that lower-performing individuals at baseline (Block 1) made more correct match decisions by Block 2. However, while this hints at an improvement in performance across blocks in the no-examples condition, similar correlations were not present for overall accuracy and mismatch trials.

By contrast, the same analysis revealed a clear negative correlation for overall accuracy, and match and mismatch trials in the examples condition. This pattern of correlations therefore indicates consistently that observers who displayed lower accuracy at baseline were more likely to improve with the provision of examples. Inspection of Figure 4 suggests the presence of three potential outliers (i.e., overall accuracy lower than 70%), but the pattern of correlations remains the same for all measures when these data points are removed (see Table 1). We also sought to explore further whether these correlations are genuinely driven by improvement in the lower-performing observers at baseline, or whether these could be attributed, at least in part, to high performing observers. This is plausible considering that the best-performing observers were at or near ceiling. Thus, these observers can essentially only maintain their baseline accuracy level in Block 2 or drop below this level, which could potentially underpin the negative correlations that were observed here. Crucially, however, such a pattern would contradict the conclusion that examples *improved* performance. To investigate this possibility, we sorted participants by their baseline accuracy and conducted a median split on these data. We then repeated the correlations

for observers with the lower and higher baseline accuracy. The outcome of this analysis is also displayed in Table 1 and confirms that the improvement correlations were driven primarily by observers with lower baseline accuracy. Taken together, the analyses indicate that examples improved face-matching accuracy, particularly in lower-performing individuals.

In a final step, we sought to explore whether the observed improvements in accuracy relate directly to individual differences in self-reported example usage. On average, participants indicated that examples were utilised on 26.6% ($SD$ = 18.7) of trials in Block 2, with individual differences ranging from 0.0 to 97.5% of trials. The individual self-report example-usage scores correlated with baseline performance (Block 1), $r$ = -.253, $p$ < .05, which suggests that observers may have possessed some awareness of the need to use examples to improve performance. However, correlation of example usage and improvement in performance (Block 2 minus Block 1) was not found, $r$ = .048, $p$ = .65. An illustration of these data is shown in Figure 5, which suggests the presence of three outliers (i.e., with baseline accuracy < 70%), but the pattern of correlations remained when these data points were removed, $r$ = -.259, $p$ < .05 and $r$ = .006, $p$ = .96, respectively. Observers were also more likely to use examples in the first half of Block 2 (31.2%) than in the second half of Block 2 (21.9%), $t(89) = 5.47$, $p$ < .001. This suggests that observers may have felt a limit to what can be learned from the examples, and so reduced their usage over time. However, accuracy was comparable for both halves of Block 2, $t(89) = 0.99$, $p$ = .33. We return to these findings in the General Discussion.

*Generalisation*

To determine additional characteristics of the examples improvement, we assessed also whether this effect persists after the examples have been removed, and whether it generalises to

other stimuli. At a group level, the examples effect observed above accounted for an overall improvement in accuracy of less than 3% from Block 1 to Block 2 (see Group-level analysis above). In terms of generalisation, only small differences in mean accuracy were observed between the no-examples and examples groups for repetitions of the target face pairs (GFMT Old: 92.5% versus 94.6%, $t(118) = 1.72$, $p = .09$) and presentation of novel face pairs from the same face set (GFMT New: 90.2% versus 92.9%, $t(58) = 1.15$, $p = .26$). Mean accuracy was also similar across the no-examples and examples groups with faces from a different stimulus set (KFMT: 64.9% versus 63.1%, $t(58) = 0.85$, $p = .40$).[1] This indicates that there was no generalisation of the examples improvement at a group level. Further generalisation analysis therefore focuses on individual differences, by investigating correlation of any accuracy improvements in these blocks with baseline performance.

For the no-examples condition, negative correlations were observed on match trials of the GFMT Old and GFMT New blocks but not for mismatch trials or overall accuracy (see Figure 6), converging with the data for Block 2 (see Individual differences section above). In contrast, for the examples condition clear negative correlations were observed for overall accuracy, and match and mismatch trials for GMFT Old face pairs, indicating that improvements in accuracy were maintained after the removal of the example stimuli. A similar pattern was observed for GFMT New faces in overall accuracy and mismatch trials, with a non-significant trend in the same direction for match trials. This indicates that the examples effect generalised to previously unseen face pairs from the same stimulus set.

A correlation for overall accuracy was also found in the examples condition with face pairs from the KFMT (see Table 2). However, in contrast to generalisation for face pairs from

[1] Full data set available as supplement for readers wishing to analyze these further.

the GFMT, this correlation was not present when match and mismatch trials were considered separately. Finally, no correlations were observed with KFMT faces in the no-examples condition.

## General Discussion

This study examined whether matching accuracy can be improved by the provision of example face pairs. Examples improved performance at group level but this effect was numerically small (2.3%) in the context of substantial individual differences in this task. At baseline, for example, accuracy varied from 50% to 100% across individuals. Considering this expected range, which converges with other studies (e.g., Bindemann et al., 2012; Burton et al., 2010; White, Kemp, Jenkins, Matheson, et al., 2014), we investigated whether any improvement with examples depends on the ability of individuals to perform this task. This revealed a consistent effect in match, mismatch and overall accuracy, whereby examples improved performance most in observers who were least accurate at the beginning of the experiment.

A similar correlation was also found for the no-examples condition, but only with match trials. At the group level an increase in match accuracy was observed across blocks, which converges with other studies (Alenezi & Bindemann, 2013; Alenezi et al., 2015; Bindemann et al., 2016; Fysh & Bindemann, 2017b; Papesh et al., 2018). In the current experiment, this effect was not affected by the presence of examples. We therefore attribute the above correlation in the no-examples condition, and the corresponding correlation in the examples condition, also to a tendency to make more match responses over the course of the experiment. There was logically more scope for lower- than higher-performing observers from the baseline block to record such

an increase in Block 2, thus leading to the negative correlations for match trials that were observed in the examples and no-examples condition.

Overall, however, we suggest that the differences between conditions in the correlational data for Block 2 indicate a clear examples effect, whereby individuals with lower baseline accuracy improve particularly with the provision of such stimuli. At a group level, this effect did not transfer to face pairs from the same stimulus set as the examples (the GFMT) or a different stimulus set (the KFMT). However, an individual differences-based generalisation effect was found for face pairs from the GFMT after the examples were removed from view, whereby observers with lower baseline accuracy demonstrated greater improvement via the correlational analysis. This was observed for faces that had been seen before as well as previously unseen GFMT faces, and was evident across all accuracy measures. For faces from the KFMT, on the other hand, such an effect was observed only in overall accuracy and was accompanied by a similar non-significant correlation in the no-examples condition, suggesting no transfer across different stimulus sets.

How might examples help to improve face-matching accuracy? The existence of individual differences in matching performance indicates that some observers simply do not know how to utilize the available facial information information to make an identification (Bindemann et al., 2012; Bobak, Dowsett et al., 2016; Bobak, Hancock et al., 2016; Bobak, Pampoulov et al., 2016; Burton et al., 2010; Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016). In turn, providing criteria about what constitutes an identity match or mismatch via feedback helps to improve performance (White, Kemp, Jenkins, & Burton, 2014). We suggest that provision of examples enhances face-matching accuracy in a similar manner.

One possibility for how this enhancement of matching criteria occurs is that the presence of examples highlights features that are shared in different photographs of the same face, and hence are diagnostic of identity matches. Similarly, the presence of mismatch examples may highlight features that distinguish different people. This explanation resonates with the diagnostic-feature-detection model of eyewitness identification (Wixted & Mickes, 2014; Wixted, Vul, Mickes, & Wilson, 2018), according to which the *simultaneous* presentation of multiple faces allows for the extraction of critical identity information about a target in a lineup that is not apparent when the same faces are presented in isolation. For example, such a comparison may highlight features that are shared across people and are therefore not diagnostic of a specific person. If these features are discounted, this may lead observers to utilise criteria that are more diagnostic of differences in identity, enhancing detection of a target.

In the current study, the examples-advantage also persisted after these were no longer present, indicating an additional effect that transcends beyond comparison of a set of concurrent faces. This indicates that the face-matching criteria that were acquired from examples were also internalised by observers. This aspect of the current findings could be explained by a decision-making framework in which the criteria required to distinguish identity matches and mismatches exhibit variability *across* trials (see Benjamin, Diaz, & Wee, 2009). Providing insight into this variability, through the presentation of labelled examples, may have served to stabilize decision criteria across trials, leading to the more lasting improvements in accuracy that were observed with GFMT faces in the third and fourth block here.

A criteria-based explanation is appealing in light of the substantial within-person variability that faces can exhibit in appearance (Jenkins, White, Van Montfort, & Burton, 2011), the seemingly idiosyncratic nature of this variability (Burton, Kramer, Ritchie, & Jenkins, 2016),

and considering that improvements in accuracy were most evident in lower-performing observers here (i.e., those who may have the least stable decision criterion). It also resonates with the lack of improvement transfer to face pairs from a different stimulus set (the KFMT; Fysh & Bindemann, 2018b) that was observed here, if the required criteria for identity matching vary across face sets. In support of this reasoning, it is already known that different visual features carry identity information in faces of different ethnicities (McDonnell, Bornstein, Laub, Mills & Dodd, 2014). If the same applies to different face sets of the same ethnicity, then this could explain why no generalisation was found for KFMT faces.

The absence of generalisation of improvement to KFMT faces is a potential limitation if one were to consider the provision of examples for improving face-matching performance in applied settings, such as passport control, where observers encounter faces from a broad range of categories. Other methods under investigation for improving face-matching accuracy, such as feature comparison strategies (Towler, White, & Kemp, 2017) and feature instructions (Megreya & Bindemann, 2018), also show limited generalisation to other stimulus sets. Thus, we cannot easily resolve this issue here. We also note that although clear examples improvement effects were observed in matching accuracy, a direct association between observers' self-reported example usage and changes in performance was not evident. This may reflect the limited insight that observers have into internal cognitive processes (see, e.g., Nisbett & Wilson, 1977, Wilson & Dunn, 2004), whereby reported example usage may not reflect *actual* usage. One way to address this limitation could be the application of eye-tracking, to objectively determine the percentage of trials on which examples are viewed. Even so, it remains to be seen whether such a measure can distinguish cases where examples are viewed but not used, from those where these are utilised by observers to enhance performance. And as we observed generalisation of the

examples improvement across several GFMT blocks in the experiment here, it is also likely that such generalisation occurred *within* blocks, which serves to further obscure correlations of example usage and improvement. Understanding more precisely when, and how, examples are used is clearly important for further progress in this field.

In conclusion, this study shows that provision of examples improves face matching, particularly in lower-performing individuals. This improvement persists after examples are removed and transfers to face pairs from the same but not to a different stimulus set. We suggest that examples aid performance by providing criteria to distinguish identity-matches and mismatches that observers otherwise have to deduce by their own judgement during face matching.

## References

Alenezi, H. M., & Bindemann, M. (2013). The effect of feedback on face-matching accuracy. *Applied Cognitive Psychology*, *27*(6), 735-753. doi: 10.1002/acp.2968

Alenezi, H. M., Bindemann, M., Fysh, M. C., & Johnston, R. A. (2015). Face matching in a long task: Enforced rest and desk-switching cannot maintain identification accuracy. *PeerJ*, *3*, e1184. doi: 10.7717/peerj.1184

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review*, *116*(1), 84-115. doi: 10.1037/a0014351

Bindemann, M., Avetisyan, M., & Blackwell, K.-A. (2010). Finding needles in haystacks: Identity mismatch frequency and facial identity verification. *Journal of Experimental Psychology: Applied*, *16*(4), 378-386. doi: 10.1037/a0021893

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. doi: 10.1037/a0029635

Bindemann, M., Fysh, M. C., Cross, K., & Watts, R. (2016). Matching faces against the clock. *I-Perception*, *7*(5), 1-18. doi: 10.1177/2041669516672219

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PLoS ONE*, *11*(2), e0148148. doi: 10.1371/journal.pone.0148148

Bobak, A. K., Hancock, P. J., & Bate, S. (2016). Super-recognisers in action: Evidence from face-matching and face memory tasks. *Applied Cognitive Psychology*, *30*(1), 81-91. doi: 10.1002/acp.3170

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of young British adults. *Frontiers in Psychology*, *7*(1378), 1-11. doi: 10.3389/fpsyg.2016.01378

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. doi: 10.1111/cogs.12231

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*(1), 286-291. doi: 10.3758/BRM.42.1.286

Dhir, V., Singh, A., Kumar, R., & Singh, G. (2010). Biometric recognition: A modern era for security. *International Journal of Engineering Science and Technology*, *2*(8), 3364-3380. Retrieved from https://www.researchgate.net/publication/50315614_BIOMETRIC_RECOGNITION_A_MODERN_ERA_FOR_SECURITY

Fysh, M. C., & Bindemann, M. (2017a). Forensic face matching: A Review. In M. Bindemann & A. M. Megreya (Eds.), *Face processing: Systems, disorders and cultural differences* (pp. 1-20). New York: Nova Science Publishing, Inc.

Fysh, M. C., & Bindemann, M. (2017b). Effects of time pressure and time passage on face-matching accuracy. *Royal Society Open Science*, *4*(6), 170249. doi: 10.1098/rsos.170249

Fysh, M. C., & Bindemann, M. (2018a). Human-computer interaction in face matching. *Cognitive Science, 42*(5), 1714-1732. doi: 10.1111/cogs.12633

Fysh, M. C., & Bindemann, M. (2018b). The Kent Face Matching Test. *British Journal of Psychology, 109*(2), 219-231. doi: 10.1111/bjop.12260

Henderson, Z., Bruce, V., & Burton, A. M. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology*, *15*(4), 445-464. doi: 10.1002/acp.718

Jenkins, R., & Burton, A. M. (2008). Limitations in facial identification: The evidence. *Justice of the Peace*, *172*, 4-6. Retrieved from http://www.visimetrics.com/docs/technical/Limitations in Facial Recognition Article.pdf

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *366*(1571), 1671-1683. doi: 10.1098/rstb.2010.0379

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313-323. doi: 10.1016/j.cognition.2011.08.001

Lander, K., Bruce, V., & Bindemann, M. (2018). Use-inspired basic research on individual differences in face identification: Implications for criminal investigation and security. *Cognitive Research: Principles and Implications*, *3*, 1-13. doi:10.1186/s41235-018-0115-6

McDonnell, G. P., Bornstein, B. H., Laub, C. E., Mills, M., & Dodd, M. D. (2014). Perceptual processes in the cross-race effect: Evidence from eyetracking. *Basic and Applied Social Psychology*, *36*(6), 478-493. doi: 10.1080/01973533.2014.958227

Megreya, A. M., & Bindemann, M. (2018). Feature instructions improve face-matching accuracy. *PLoS ONE, 13*(3), e0193455. doi: 10.1371/journal.pone.0193455

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865-876. doi: 10.3758/BF03193433

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*(4), 364-372. doi: 10.1037/a0013464

Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching face images taken on the same day or months apart: The limitations of photo ID. *Applied Cognitive Psychology*, *27*(6), 700-706. doi: 10.1002/acp.2965

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231-259. doi: 10.1037/0033-295X.84.3.231

Özbek, M., & Bindemann, M. (2011). Exploring the time course of face matching: Temporal constraints impair unfamiliar face identification under temporally unconstrained viewing. *Vision Research*, *51*(19), 2145-2155. doi: 10.1016/j.visres.2011.08.009

Papesh, M. H. (2018). Photo ID verification remains challenging despite years of practice. *Cognitive Research: Principles and Implications, 3*:19. doi: 10.1186/s41235-018-0110-y

Papesh, M. H., & Goldinger, S. D. (2014). Infrequent identity mismatches are frequently undetected. *Attention, Perception, & Psychophysics*, *76*(5), 1335-1349. doi: 10.3758/s13414-014-0630-6

Papesh, M. H., Heisick, L. L., & Warner, K. M. (2018). The low-prevalence effect in unfamiliar face-matching: The roles of feedback and criterion shifting. *Journal of Experimental Psychology: Applied, 24*(3), 416-430. doi: 10.1037/xap0000156

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8-13. doi: 10.1016/j.jneumeth.2006.11.017

Robertson, D. J., Middleton, R., & Burton, A. M. (2015). From policing to passport control: The limitations of photo ID. *Keesing Journal of Documents and Identity*, *February*, 3-8. Retrieved from https://www.researchgate.net/profile/David_Robertson31/publication/305429407_From_policing_to_passport_control_The_limitations_of_photo_ID/links/578e690c08aecbca4cad01a.pdf

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by metropolitan police super-recognisers. *PLoS ONE*, *11*(2), e0150036. doi: 10.1371/journal.pone.0150036

Towler, A., White, D., & Kemp, R. I. (2014).

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. doi: 10.1371/journal.pone.0211037

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. doi: 10.1037/xap0000108

White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin & Review*, *21*(1), 100-106. doi: 10.3758/s13423-013-0475-3

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PLoS ONE*, *9*(8), e103510. doi: 10.1371/journal.pone.0103510

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*, 1814-1822. doi: 10.1098/rspb.2015.1292

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology*, *55*, 493-518. doi: 10.1146/annurev.psych.55.090902.141954

Wirth, B. E., & Carbon, C.-C. (2017). An easy game for frauds? Effects of professional experience and time pressure on passport-matching performance. *Journal of Experimental Psychology: Applied*, *23*(2), 138-157. doi: 10.1037/xap0000114

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, *121*(2), 262-276. doi: 10.1037/a0035940

Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. M. (2018). Models of lineup memory. *Cognitive Psychology*, *105*, 81-114. doi: 10.1016/j.cogpsych.2018.06.001

TABLE 1. Summary of Accuracy Correlations of Baseline Performance with Improvement from Block 1 to Block 2, for the Examples and No-Examples Conditions for Overall, Match and Mismatch Accuracy. Correlations are Provided for All Data and Without Potential Outliers (*All* and *W/o Outliers*), and for Median-Split Data to Show Correlations for the Worst (*Lower Accuracy*) and Best Performers (*Higher Accuracy*) at Baseline. * $p < .05$, ** $p < .01$, *** $p < .001$

| | | Change in Performance | | | |
|---|---|---|---|---|---|
| | | *All* | *W/o outliers* | *Lower accuracy* | *Higher accuracy* |
| *No-Examples* | Overall | r = -.207 | r = -.207 | r = .003 | r = -.270 |
| | Matches | r = -.421 *** | r = -.421 *** | r = -.400* | r = -.456 ** |
| | Mismatches | r = -.105 | r = -.105 | r = .010 | r = -.047 |
| *Examples* | Overall | r = -.598 *** | r = -.526 *** | r = -.453 ** | r = .059 |
| | Matches | r = -.784 *** | r = -.855 *** | r = -.751 *** | r = -.482 ** |
| | Mismatches | r = -.385 *** | r = -.330 ** | r = -.302 * | r = .086 |

TABLE 2. Summary of Accuracy Correlations of Baseline Performance with Improvement When Previously Seen Faces from the GFMT are Repeated Without Examples (GFMT Old), Previously Unseen Faces from the GFMT are Shown (GFMT New), and KFMT Faces are Shown. * $p < .05$, ** $p < .01$, *** $p < .001$

| | | Change in Performance | | |
|---|---|---|---|---|
| | | *Block GFMT Old* | *Block GFMT New* | *Block KFMT* |
| *No-Examples* | Overall | r = -.153 | r = -.007 | r = -.285 |
| | Matches | r = -.841 *** | r = -.789 *** | r = -.051 |
| | Mismatches | r = .119 | r = .325 | r = .062 |
| *Examples* | Overall | r = -.636 *** | r = -.447 * | r = -.398 * |
| | Matches | r = -.783 *** | r = -.343 | r = .026 |
| | Mismatches | r = -.639 *** | r = -.555 ** | r = -.205 |

FIGURE 1. Illustration of a stimulus of the experimental condition, comprising a centrally-presented pair of target faces, and labelled example match and mismatch pairs. In the no-examples condition, the match and mismatch stimuli to the left and right of the target pair were not shown.

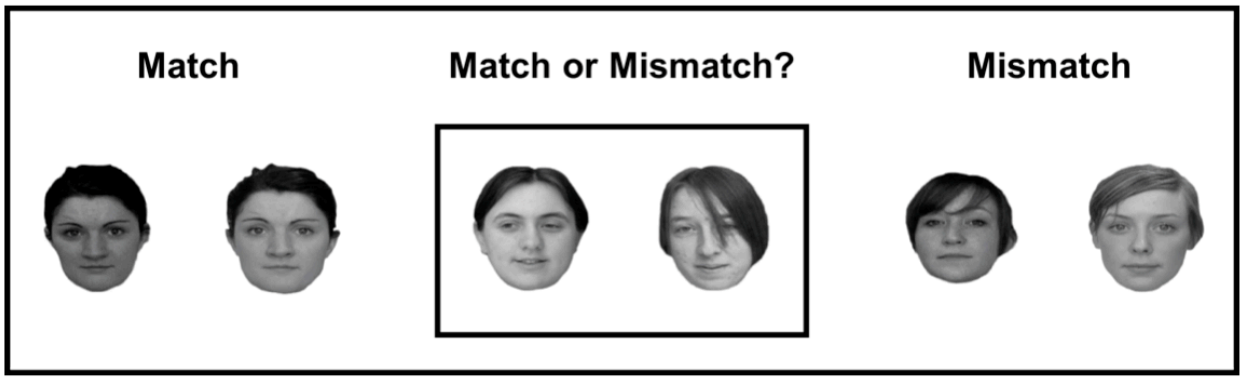

FIGURE 2. Illustration of the procedure for participants including the subset who completed a further block of the original GFMT stimuli (GFMT Old) after Block 2, as well as either a block of previous unseen GFMT stimuli (GFMT New) *or* a block of faces taken from a different stimulus set (KFMT).

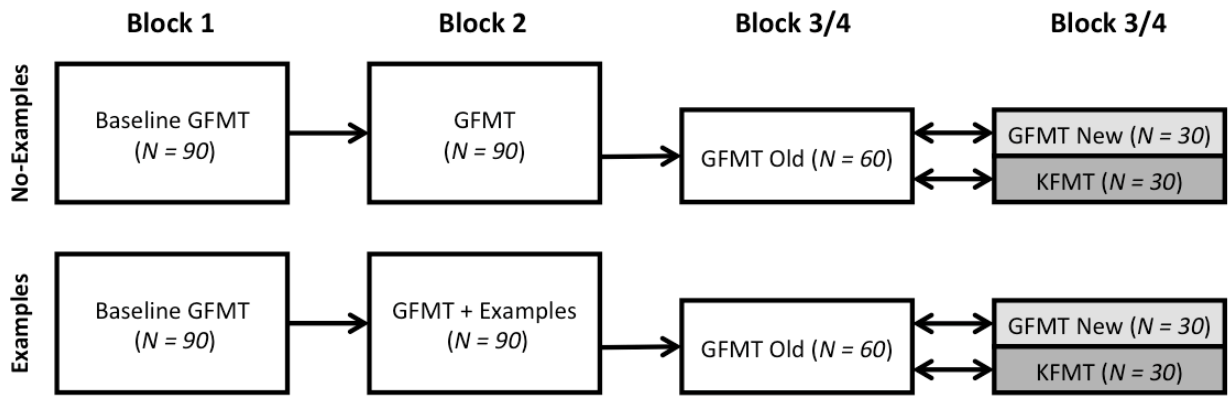

FIGURE 3. Mean percentage accuracy for the examples condition (top left) and the no-examples conditions (top right) as a function of block and trial type, and an illustration of the block by condition interaction (bottom left) and the block by trial type interaction (bottom right). Error bars show standard error of the means.

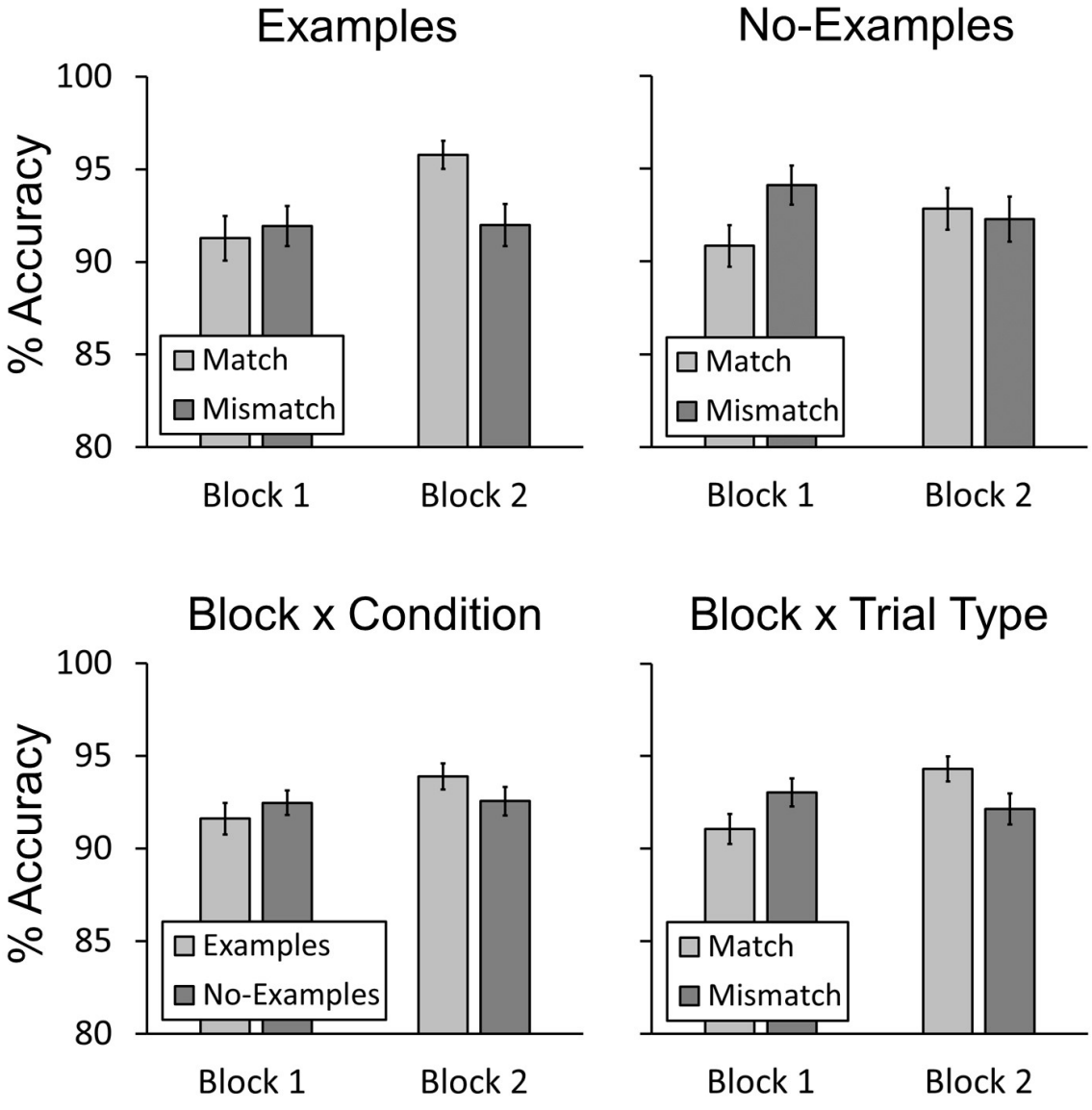

FIGURE 4. Baseline accuracy correlated with improvement from Block 1 to Block 2. Black markers denote potential outliers (observers who scored less than 70% overall at baseline).

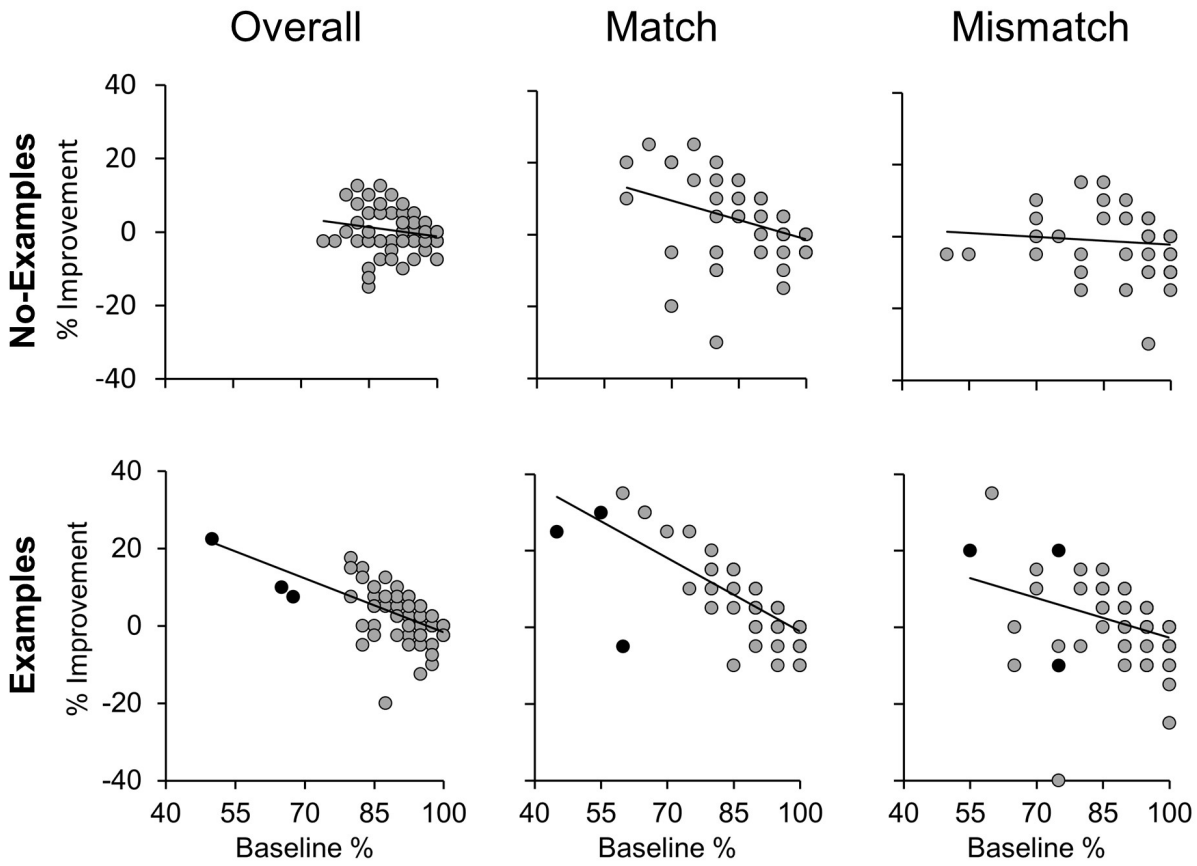

FIGURE 5. Example usage correlation with baseline (Block 1) accuracy and change in performance (from Block 1 to Block 2).

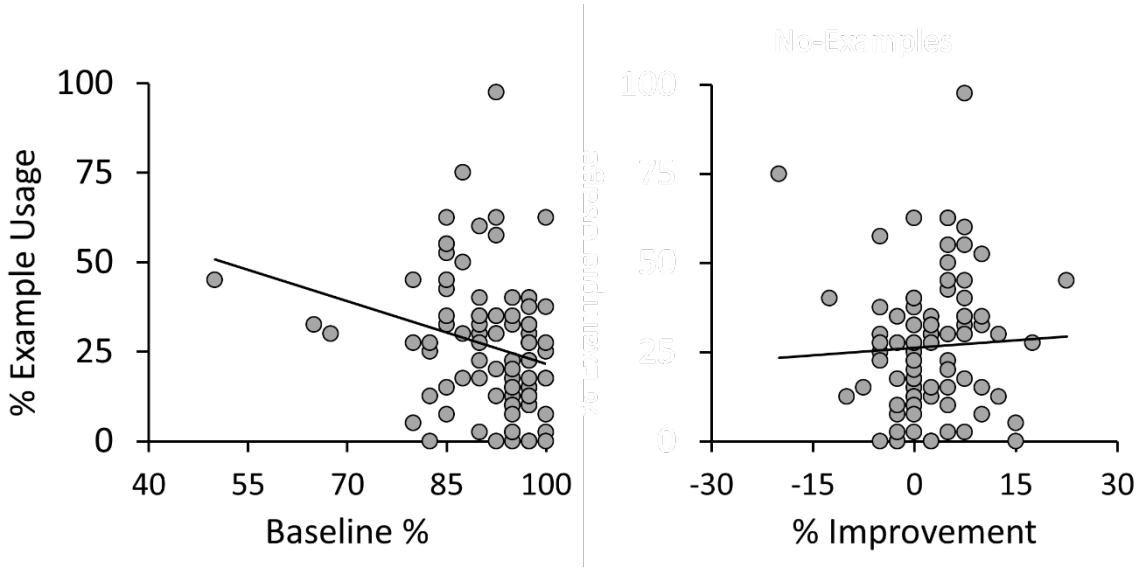

FIGURE 6. Correlations of baseline accuracy and improvement from Block 1 to Block GFMT Old, Block GFMT New, and Block KFMT for the Examples and No-Examples conditions.