Metacognition during unfamiliar face matching

Robin S. S. Kramer[*1], Georgina Gous[1], Michael O. Mireku[1], and Robert Ward[2]

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

[2] School of Psychology, Bangor University, UK

[*]Corresponding author information: Robin S. S. Kramer, School of Psychology, University of Lincoln, Lincoln, LN6 7TS, UK (email: remarknibor@gmail.com).

Data availability statement:

The data that support the findings of this study are openly available in Open Science Framework at https://osf.io/gew49/?view_only=bf1a9717d19e40fbbd28b4a685941699

Acknowledgements:

The authors thank our Research Skills III and IV students for collecting the data, as well as Abi Davis for providing critical comments on the manuscript.

Abstract

Kruger and Dunning (1999) described a metacognitive bias in which insight into performance is linked to competence: poorer performers are less aware of their mistakes than better performers. Competence-based insight has been argued to apply generally across task domains, including a recent report investigating social cognition using a variety of face-matching tasks. Problematically, serious statistical and methodological criticisms have been directed against the traditional method of analysis used by researchers in this field. Here, we further illustrate these issues and investigate new sources of insight within unfamiliar face matching. Over two experiments (total $N = 1077$), where Experiment 2 was a preregistered replication of the key findings from Experiment 1, we found that insight into performance was multi-faceted. Participants demonstrated insight which was *not* based on competence, in the form of accurate updating of estimated performance. We also found evidence of insight which *was* based on competence: the difference in confidence on correct versus incorrect trials increased with competence. By providing ways that we can move beyond problematic, traditional approaches, we have begun to reveal a more realistic story regarding the nature of insight into face perception.

Keywords

Metacognition; Face matching; Dunning-Kruger effect; Insight; Confidence

In a ground-breaking series of experiments, Kruger and Dunning (1999) investigated the self-insight of problem-solvers across a variety of domains. Figure 1 is taken from Experiment 1 of Kruger and Dunning, and is indicative of their own and also many subsequent experiments using their methods. The figure plots self-estimates of performance as a function of actual performance quartile, or more conceptually, self-insight as a function of competence.

There is an imperfect correlation between self-insight and competence in Figure 1. We will later discuss some of the difficulties of interpretation this raises, but for now we focus on the two main conclusions drawn by Kruger and Dunning (1999) from this pattern of data. First, we can see that the least competent quartile shows the least self-insight, that is, the greatest discrepancy between estimated and actual performance. The conclusion drawn by Kruger and Dunning is that one's metacognitive ability, or insight into one's own thought processes, depends upon one's competence. We will call this *competence-based insight*. Second, from Figure 1 we can also see the lack of insight found in low-competence quartiles was manifest as *overconfidence*, such that the self-estimated performance of those in the lower quartiles was higher than their actual performance.

The pattern shown in Figure 1[1], and since replicated across a variety of domains (e.g., humour – Kruger & Dunning, 1999; political knowledge – Anson, 2018; wine knowledge – Aqueveque, 2018; reasoning – Pennycook et al., 2017; pilot knowledge – Pavel et al., 2012), led Kruger and Dunning (1999) to argue that the skills needed to perform well in a domain will often be the same skills needed to evaluate performance in that domain. Poor performers will therefore be less able to assess their true level of performance – a failure of insight. As an extreme characterisation, someone who is unable to perceive differences in musical pitch would have low performance in a pitch-matching task but also would be unable to assess how

[1] Note also that the highest quartile slightly underestimated their performance. The explanation for this underestimate is frequently couched in terms of the false-consensus effect (Ross et al., 1977). By this false consensus, the highest-performing individuals are assumed to realise they have performed well, but mistakenly assume that others would also provide the same (i.e., correct) answers as they themselves have.

well they were matching. Kruger and Dunning went on to claim that competence-based insight meant that poor performers were under a "dual burden" – they do not perform well and yet are not in a position to recognise how to improve.

Although competence-based insight and overconfidence from the lowest performers seem to co-occur in Figure 1, these two phenomena are dissociable in theory. Competence-based insight could in fact be manifest in the form of under-confidence – for example, low competence performers might have essentially zero confidence in all answers, and so again be unable to recognise and correct their errors. This example shows that competence-based insight is not a function of overall confidence about a task, but in differential confidence – whether confidence is indicative of whether one's answer is correct or not. We emphasise that regardless of any associated effect with confidence, "competence-based insight" identifies Kruger and Dunning's (1999) "dual burden": low competence is associated with low insight.

Competence-based insight and confidence could also have different real-world implications. For example, the effects of being unable to recognise and correct errors might be very different depending upon whether this was associated with over- or underestimates of actual performance. It is plausible that overconfident but poor performers are a greater risk in many domains, for example, driving heavy machinery. But we could reasonably speculate that unwarranted under-confidence might also be harmful, perhaps in cases where actions that should be taken are not.

The above considerations simply show it is reasonable to keep competence-based insight as a separate concept from over- and under-confidence. However, it is often the case that when research refers to the "Dunning-Kruger Effect" (DKE), it is not clear whether it is referring to competence-based insight, overconfidence by the lower performers, or both. Given that these two aspects of performance are theoretically dissociable and have different

implications, here we will try to keep them conceptually separate. We will use the term *metacognition* when we need to refer to insight and confidence together. However, our main focus will be on competence-based insight, as the ultimate basis for the "dual burden" identified by Kruger and Dunning (1999).

Face processing and possibilities for dissociable social metacognition

Although competence-based insight appears to hold across many cognitive domains, it is an important and theoretically interesting question whether this is true for social tasks in particular. Metacognition for social situations and episodes is frequently framed in terms of the function of a "theory of mind", our human ability to understand and reason about other people's beliefs and states (e.g., Baron-Cohen et al., 1985; Kuhn, 2000), and which is thought to be mediated by specific brain networks (e.g., Richardson & Saxe, 2020). It is therefore plausible that metacognition about social processes may operate differently from the general pattern found in non-social processes. Face processing is particularly interesting in this respect in that it relies upon highly specialised brain regions for competent performance (Haxby et al., 2000), and people are often surprised at how difficult face processing tasks can be (e.g., Kramer et al., 2019, 2020). Therefore, if we are seeking to find *domain-specific* metacognition, face processing seems like a good place to start: a process that might involve a specialised theory of mind network and that frequently leaves people surprised at their true competence.

Furthermore, the accuracy of metacognition associated with face processing raises practical issues for assessing face processing ability in a wider societal context. For example, law enforcement agencies worldwide are expected to determine whether unfamiliar people are indeed who they claim to be, typically through the use of identification document matching (e.g., border force officers). In these settings, it is crucial that the best performers

are selected and utilised, and such individuals show good insight into their own abilities. Both under- and over-confidence in such settings could have serious repercussions.

Previous research has found only moderate insight into one's own face perception abilities, as assessed by the simple association between competence and questionnaire-based self-estimates. For example, the 20-item prosopagnosia index (Shah et al., 2015), which was developed to quantify prosopagnosic traits, has demonstrated medium-sized associations in a number of studies (Gray et al., 2017; Livingston & Shah, 2018; Shah et al., 2015; Ventura et al., 2018) with performance on both the Glasgow Face Matching Test (GFMT; Burton et al., 2010) and the Cambridge Face Memory Test (Duchaine & Nakayama, 2006). Other self-report measures have also shown similar medium-sized associations with additional matching and memory tasks (Bobak et al., 2019; Kramer, 2021; Matsuyoshi & Watanabe, 2021), although not with emotion recognition (Kelly & Metcalfe, 2011). Indeed, even when estimates are in relation to the task itself (e.g., "how will I perform on this matching test?") rather than some general face perception ability, the correlation remains only moderate (.27 for unfamiliar face matching; Zhou & Jenkins 2020).

The possibility that face perception might be a form of social cognition that is an exception to competence-based insight was explicitly tested by Zhou and Jenkins (2020). In a series of studies, these researchers investigated several types of face matching (identity, gaze direction, emotional expression), and in all cases, reported metacognitive errors. Specifically, the lowest performers overestimated, and the highest underestimated, their performance. These results suggest that, despite the special nature of face perception, it is subject to the same metacognitive errors as other domains. However, these studies used the procedures and analyses illustrated in Figure 1, and so these conclusions rest on the interpretation of an imperfect correlation between actual and estimated performance.

Issues with DKE methodology

Although the questions addressed by Zhou and Jenkins (2020) are clearly important, we now turn to significant complications with the methods and analyses used. Zhou and Jenkins followed the strategy employed by Kruger and Dunning (1999; Experiments 1, 2, and 3): they separated their sample into performance-based quartiles, and compared actual to estimated performance. However, serious statistical and methodological criticisms have been directed against this specific approach to assessing metacognition (e.g., Gignac & Zajenkowski, 2020; Krajc & Ortmann, 2008; Krueger & Mueller, 2002; Meeran et al., 2016; Nuhfer et al., 2016, 2017).

First, it has long been understood that the use of quartiles in this context is both potentially problematic and completely unnecessary. Dividing a distribution of continuous performance into arbitrary categories can misrepresent the true pattern of data, increase the risk of false positives, and lower the power of the experiment to measure the true correlation of estimated and actual performance (Altman & Royston, 2006; MacCallum et al., 2002; McClelland et al., 2015).

Second and more significant, the conclusions that can be drawn from an imperfect correlation between actual and estimated performance are limited, and in fact, this imperfect correlation is almost uninformative about metacognitive processes, due to a combination of factors. First, on purely statistical grounds, even if all quartiles had perfect insight into their abilities, regression towards the mean will produce a flattened slope of estimated compared to actual performance. Even if every individual had perfect knowledge of their actual competence, and how they would perform over a large number of tests, the score on any given test represents a combination of both actual competence and random chance: the best performances will have had some unpredictable good luck and the worst some unpredictable

bad luck (e.g., Kahneman et al., 2021). The imperfect correlation might therefore not indicate competence-based insight, but simply that actual competence does not completely predict the combination of actual competence plus random chance[2].

Further, in the example shown in Figure 1, the lower quartiles are the least accurate overall in judging their abilities, in that they show the largest difference between actual and estimated scores. This might seem at first glance to suggest some lack of metacognition in these quartiles. However, the quartile that is expected to show the largest gap between estimated and actual scores (that is, nominally, the least insight) depends upon the intercept of the function describing estimated and actual performance. This intercept, in turn, has been shown to depend upon perceived task difficulty (Burson et al., 2006). For example, when a task is perceived as very difficult, estimated performance for all groups drops, such that the entire sample may be underestimating their actual performance. In this case, it is the highest quartiles that can have the largest gap between estimated and actual performance. The effects of perceived task difficulty mean that the classic pattern seen in Figure 1 is expected in circumstances where participants estimate the difficulty of the task (and therefore the vertical placement of the 'perceived ability' line) as somewhere between the scores obtained by the lowest and highest quartiles. Finally, the intercept may also be influenced by the better-than-average effect (Mabe & West, 1982), which tends to lower perceived task difficulty. This causes the intercept of the estimated and actual performance slope to be relatively high on the *y*-axis (since the 'perceived ability' line is generally higher). This results in the average level of underestimation (for the highest scorers) being less than the average level of overestimation (for the lowest scorers). Taken together, these mechanisms – which have

[2] This was acknowledged by Dunning (2011), who suggested some statistical approaches to try to compensate for a 'regression to the mean' effect.

arguably nothing to do with insight – are expected to produce, under a wide variety of circumstances, precisely the pattern reported by Kruger and Dunning (1999).

However, perhaps the most compelling argument against insight-based interpretations of data like those illustrated in Figure 1 is that results like these can be reproduced with entirely randomised data (Nuhfer et al., 2016, 2017). In randomly generated datasets of sufficient size, the mean estimated performance for all four quartiles will be approximately equal. As discussed above, any estimation slope that is less steep than the accompanying slope of actual scores will result in an apparent competence-based insight.

We verified that the arguments raised by Nuhfer et al. (2016, 2017) are a challenge for the Zhou and Jenkins (2020) dataset. Using the summary data made available online by Zhou and Jenkins, and focussing on their unfamiliar face matching experiment, we investigated what would happen if estimated performance scores were randomly shuffled. That is, what pattern of results would we see if there was no relationship whatsoever between actual and estimated scores? Figure 2a reproduces the original pattern of results (see Figure 2 in Zhou & Jenkins, 2020) whereas Figure 2b illustrates five iterations of shuffled data.

Unfortunately, as Figure 2 illustrates, flattened slopes with overestimates by lowest performers and underestimates by highest performers appear even when the estimated performance estimates are randomly shuffled among participants. This pattern used by Zhou and Jenkins (2020) is therefore undiagnostic, as it can easily arise as a statistical artefact of the analysis process (Nuhfer et al., 2016, 2017).

Finally, the estimated performance measure used by Zhou and Jenkins (2020) is somewhat difficult to interpret in the context of metacognitive ability. This is because Zhou and Jenkins did not report actual participant estimates of task performance, but a derived estimate of those estimates that assumes important metacognitive abilities. After each response, participants indicated whether they were “sure” or “unsure” of its accuracy. The

frequency of trial-level "sure" responses for each participant was then converted into that participant's estimated performance score for the entire task. However, this conversion included a guessing correction, by adding half of the frequency of "unsure" trials to the estimate, in order to correct for guessing on the two-alternative forced choice task used. For example, 24 "sure" responses out of 40 trials would result in a derived task-level estimate of 32 out of 40. Corrections due to guessing are a clear form of reasoning about true performance, i.e., metacognition. Since metacognition is the trait we are seeking to understand, it seems preferable to avoid making assumptions about metacognition in this calculation. In addition, it is worth noting that this method conflates task- and trial-level insight (by using the former to derive an estimate of the latter), which previous research has shown to be dissociable (e.g., Kelly & Metcalfe, 2011).

Competence and trial-by-trial confidence

The above considerations mean that we should approach the issue of competence-based insight in face processing (or indeed in any process) in a different way to that of Kruger and Dunning (1999; Experiments 1, 2, and 3), and Zhou and Jenkins (2020). We sought to employ what should be more robust ways to assess competence-based insight, and in particular, we wanted to avoid the comparison of globally estimated versus actual accuracy (e.g., Figure 1) and its associated issues: regression to the mean, the influence of perceived task difficulty on interpretation of over- and underestimates of performance, and the possibility that even randomised data can produce the pattern of interest. We instead focussed on item-by-item measures of confidence, and the influence of actual performance (i.e., competence) on whether confidence was diagnostic of accuracy.

Our reasoning is as follows. First, at the group level, confidence ratings are associated with response accuracies during face perception tasks. That is, the mean confidence rating for

trials answered correctly is higher than for trials receiving incorrect responses (Bruce et al., 1999; Hopkins & Lyle, 2020; Stephens et al., 2017). At this group level of analysis, confidence is a diagnostic cue for correct compared to incorrect responses for face processing tasks, including face recognition (Grabman & Dodson, 2020), searching for faces in crowds (Davis et al., 2018; Kramer et al., 2020), and identifying faces that were present in previously shown arrays (Ji & Hayward, 2020).

These results show that trial-level confidence responses can reflect insight about performance, at a group level. But the second and crucial consideration relates to Kruger and Dunning's (1999) "dual burden": is insight based on competence? Do poorer performers show a reduced difference in confidence on correct versus incorrect trials compared to better performers? Kelly and Metcalfe (2011) collected participants' ratings of confidence after each response during tasks designed to assess emotional expression recognition. While participants in both tasks showed evidence of trial-level insight (higher confidence ratings for correct responses), only one task provided support for its relation with competence – those who were better at the emotion recognition task also tended to show greater insight, that is, a larger difference in confidence for correct versus incorrect responses. In addition, using a test of mathematical knowledge, Händel and Dresel (2018) found that the least competent performers were more confident for incorrect in comparison with correct responses, while the highest performers were more confident when their responses were correct. To date, no research has considered this relationship with regard to face matching abilities.

The current study

In the current experiments, we assessed competence-based insight in multiple ways. We asked participants for their estimated performance, in both absolute (how well did you do?) and relative terms (how well did others do?), and both before and after the task. Crucially, we

also asked how confident each participant was in their response for every trial. Together, these measures give us multiple ways to assess insight:

(1) *The difference in confidence for trials the participant answered correctly versus incorrectly.* Insight into performance is indicated by greater confidence on correct as compared with incorrect trials. If competence is associated with insight, then we expect the difference in confidence for correct and incorrect trials to be correlated with overall task accuracy.

(2) *The difference between pre- and post-task estimates of performance*. The initial estimate of participants on how well they will do on a task, before actually performing that task, could be incorrect for many reasons. However, if a participant has insight into their performance, then the post-task estimate of performance should be closer to reality than the pre-task estimate. As a result, we would expect updating of performance estimates across our sample to vary according to how erroneous participants' initial estimates were of their performance.

(3) *Insight into relative performance*. By considering the difference between participants' estimates of their own and other people's performance, we can measure insight into performance relative to others. For there to be evidence of insight in our data, we would expect this difference to correlate with overall task performance.

By identifying separable measures of competence-based insight, we may find evidence of insight in face processing that is not the product of particular statistical approaches.

Experiment 1

This first experiment was exploratory in nature. Our aims were to compare the different insight measures above in a well-powered sample.

Method

Participants

A representative sample of 614 volunteers (396 women; age $M = 26.7$ years, $SD = 12.5$ years; 94% self-reported ethnicity as White) gave informed, onscreen consent before participating in the experiment and were provided with an onscreen debriefing upon completion. Participants were recruited by word of mouth (e.g., through asking friends and family, and sharing the experiment's weblink on social media). Both experiments reported here were approved by the university's School of Psychology ethics committee (PSY2021002) and were carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki. There was no overlap between this sample and those who participated in Experiment 2.

The data from one additional participant were excluded because their competence (32.5%) on the GFMT (Burton et al., 2010) was substantially below 50%, which represented chance-level performance. Indeed, none of the participants in the original study scored less than 51% on this test (Burton et al., 2010), providing further justification for discarding these data.

An *a priori* power analysis was conducted using G*Power 3.1 (Faul et al., 2007), based on the effect size ($\eta_p^2 = 0.30$) of the previously reported interaction between Measure and Quartile for unfamiliar face matching (Zhou & Jenkins, 2020). In order to achieve 95% power at an alpha of .05, a total sample size of 20 was required. However, we utilised an oversampling strategy here in order to gain a larger dataset for investigation, given the exploratory nature of this first experiment.

Stimuli

We used the short version of the GFMT (Burton et al., 2010) in order to assess face matching performance. The task comprised 40 pairs of adult male (24) and female faces (16), where half the pairs were match trials (different images of the same person, taken approximately 15 min apart with different cameras) 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 (see Figure 3). The 40 face pairings were taken from the original GFMT set of 168 pairs and represented the most difficult trials (based on the performance of 300 participants; Burton et al., 2010).

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).

Participants were first provided with information regarding the test they were about to complete: pairs of face photographs would be presented; the people in these photographs would be unfamiliar to them; and that each pair of photographs would show either the same person (the two photos were taken with different cameras) or two different people (who were chosen to look similar in appearance). For each pair, participants were told that they must decide whether the photographs showed the same person or two different people.

In addition, approximately half of the participants (allocated randomly) were shown an example 'match' trial and 'mismatch' trial (shown in Figure 3), labelled as such, in order to provide additional information about the test. No examples were shown to the remaining participants. Given that the short version of the GFMT comprised the most difficult trials from the original GFMT, these example pairs were selected from the original version and represented the most difficult match and mismatch trials that: a) did not appear in the short

version of the test; and b) did not feature any identities that appeared in the short version of the test. As such, these two examples were representative of the test's difficulty while not providing any information regarding particular test trials.

Next, participants were asked, "How well do you think you will do on the test? (40 questions; each has two responses to choose from.)" Responses were given using an onscreen slider with endpoints labelled as 20 (chance performance, only guessing) and 40 (perfect score). Participants were then asked, "What do you expect the average score to be on this test for everyone else who takes part?", with responses again provided using this slider.

After answering these two questions, participants completed all 40 trials of the short version of the GFMT. 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. In addition, participants were asked "how confident are you in your response?", providing a rating from 0 (not at all confident) to 5 (extremely confident). 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 GFMT, participants answered two final questions: "How well do you think you did on the test? (40 questions; each has two responses to choose from.)" and "What do you expect the average score to be on this test for everyone else who takes part?" Responses to both questions utilised the same slider described above.

Results

Overall performance on the GFMT (M = 87.72%, SD = 9.52%, range = 52.5%-100%) was comparable with levels found in previous studies (81% – Burton et al., 2010; 86% – Kramer et al., 2020; 85% – Kramer & Reynolds, 2018).

Although half of the participants were provided with example items before completing the test, we found no evidence that this manipulation affected responses (see the supplementary materials). We therefore collapsed our data across these two subsamples in the analyses that follow. One likely explanation for the lack of an effect might be that participants simply paid little attention to the examples when presented, although this could not be confirmed here.

Traditional Dunning-Kruger analysis using quartiles

Despite its significant issues, for easier comparison to past research, we followed the traditional analysis used by Kruger and Dunning (1999), Zhou and Jenkins (2020), and numerous others (e.g., Dunning et al., 2003). We separated our participants into performance quartiles using their actual GFMT test scores. Participants' estimates of their own performance were then averaged within each quartile and compared to actual performance. Analysing estimates produced before completing the test, we carried out a 2 (Measure: actual score, estimated score) x 4 (Quartile: lowest, second, third, highest) mixed ANOVA, with Measure varying within participants while Quartile varied between participants. The main effect of Measure was statistically significant, $F(1, 610) = 666.50$, $p < .001$, $\eta_p^2 = 0.52$, as was the main effect of Quartile, $F(3, 610) = 159.37$, $p < .001$, $\eta_p^2 = 0.44$. However, these were qualified by a significant interaction between the two factors, $F(3, 610) = 89.28$, $p < .001$, $\eta_p^2 = 0.31$. We therefore considered the simple main effects of Measure at each level of Quartile. These simple main effects were statistically significant for the second, $F(1, 610) = 169.53$, $p < .001$, $\eta_p^2 = 0.22$, third, $F(1, 610) = 270.60$, $p < .001$, $\eta_p^2 = 0.31$, and highest quartiles, $F(1, 610) = 402.01$, $p < .001$, $\eta_p^2 = 0.40$, but not for the lowest quartile, $F(1, 610) = 0.17$, $p = .680$, $\eta_p^2 = 0.00$.

We also carried out the same analysis for estimates of participants' own performance produced *after* completing the test. The main effect of Measure was statistically significant, $F(1, 610) = 858.93$, $p < .001$, $\eta_p^2 = 0.59$, as was the main effect of Quartile, $F(3, 610) = 234.80$, $p < .001$, $\eta_p^2 = 0.54$. Again, these were qualified by a significant interaction between the two factors, $F(3, 610) = 38.41$, $p < .001$, $\eta_p^2 = 0.16$. Simple main effects were statistically significant for all quartiles: lowest – $F(1, 610) = 44.23$, $p < .001$, $\eta_p^2 = 0.07$; second – $F(1, 610) = 264.09$, $p < .001$, $\eta_p^2 = 0.30$; third – $F(1, 610) = 287.73$, $p < .001$, $\eta_p^2 = 0.32$; highest – $F(1, 610) = 322.66$, $p < .001$, $\eta_p^2 = 0.35$. The results of both analyses are combined for illustrative purposes in Figure 4.

Whether these results demonstrate the classic DKE pattern (e.g., as illustrated in Figure 1) or not depends upon one's interpretive stance. In absolute terms, it is the lowest performers who are most accurate in their estimates. However, as previously shown by Burson et al. (2006) and as we discussed earlier, if the intercept of these estimated lines had been 15 points higher (i.e., if the test had been perceived as easier), then we would have the apparent result that low performers overestimate and high performers underestimate. That is, as discussed previously, the classic approach gives no satisfactory answer as to whether one group has more insight than another. All we really know is that there is an imperfect correlation of actual and estimated scores, and this is hardly surprising. Further, as can be seen in the supplementary materials (Figure S1), the same pattern can be produced by simply shuffling the participants' estimates of their performance (Nuhfer et al., 2016, 2017). This simply reinforces the claim we made earlier that the "classic" analysis technique of Kruger and Dunning (1999), as well as Zhou and Jenkins (2020), is undiagnostic about competence-based insight.

Estimating own performance in relation to others

Participants estimated their own performance, and also how well they thought others would perform on the test, both before and after completing the GFMT. Insight would be demonstrated by accurately assessing one's performance relative to the group, even if performance estimates were, in absolute terms, too high or too low. (Absolute accuracy would depend on whether participants could accurately determine the difficulty of the test.) For each participant, we calculated the difference between the participant's estimate of their own performance and that of other people, separately for responses given before and after completing the test ('own_estimate – others_estimate'). As such, positive values represented those who thought they had performed 'better than average'. We then investigated the association between this variable and participants' actual performance on the test.

For estimates given *before* the test, we found a nonsignificant correlation with actual performance, of trivial effect size, $r(612) = .07$, $p = .069$. For estimates provided *after* completing the test, the association was small to moderate and significant, $r(612) = .21$, $p < .001$. A comparison of these correlations showed a statistically significant difference, $z = 2.48$, $p = .013$. As such, before undertaking the GFMT task, participants were largely unaware of how they would do relative to others, regardless of their actual ability. However, upon completion of the task, participants as a group made accurate estimates of their relative performance. This result demonstrates insight across our sample, that is, insight that is not competence-based.

Insight into performance after the test

Next, we investigated participants' updating of their own estimated abilities on the GFMT by focussing on how their estimated scores differed from their actual scores. For each participant, we calculated the difference between their estimate of their own score, given before completing the test, and their actual test score (Estimation Error = actual_score –

own_estimate_before). We also calculated the difference between their estimates of their own score given before and after the test (Estimation Updating = own_estimate_after – own_estimate_before). We found a large correlation between these two computed measures, $r(612) = .55$, $p < .001$ (see Figure 5). This association also remained when we considered each quartile separately (lowest: $r = .55$; second: $r = .42$; third: $r = .64$; highest: $r = .53$). Therefore, the greater the difference between participants' estimates given beforehand and their actual performance on the test, the larger the subsequent change in their 'before' versus 'after' estimates of themselves. For instance, if their actual score on the test was better than the estimate they gave beforehand, their 'after' estimate of performance would tend to be higher than their 'before' estimate. Equally, if they performed worse than they had estimated beforehand, they would lower their 'after' estimate relative to their 'before' estimate. This updating occurs in the absence of explicit feedback, and is based on participants' own insights into their performance. Again, by this measure, participants across our sample showed insight into their performance that was not competence-based.

Confidence and competence-based insight

In addition to insights at the level of overall test performance, we employed an analysis to investigate whether participants showed trial-level insight into their abilities, in terms of how confident they were in their responses. Insight would be demonstrated when a participant is more confident in their correct responses than their incorrect ones. To investigate how competence might affect such differences in confidence, the trial-level data were analysed using linear mixed-effects models. We used crossed random effects (participants and trials) because each participant completed the same series of trials. Therefore, participants and trials variance were considered at Level 2 and residual variance at Level 1. In terms of the dataset, each participant by trial observation was the unit of analysis, with each row of data indicating

the participant's overall score on the test (Competence), the confidence rating given by the participant to that trial and whether the response given was correct or incorrect (Accuracy).

The fixed effects were the intercept and the effects of Competence and Accuracy. In this model, only the intercept varied randomly across trials, whereas the intercept and the slope of the Accuracy varied randomly across participants. Models using more complex random effects structures were identified as singular (Barr et al., 2013). Statistical analyses were carried out using R (lme4 package – Bates et al., 2015). For significance reports, degrees of freedom were estimated using Satterthwaite's method (lmerTest package – Kuznetsova et al., 2017).

First, we fitted a model in which the interaction between Competence and Accuracy was not included. We found a significant main effect of Accuracy, ß = 0.53, *SE* = 0.02, $t(461) = 21.43$, $p < .001$, such that correct responses were given higher confidence ratings. We also found a significant main effect of Competence, ß = 0.01, *SE* < 0.01, $t(617) = 4.39$, $p < .001$, such that more competent participants gave higher confidence ratings.

Next, we included the crucial Competence x Accuracy interaction to allow for the possibility that participants' abilities influenced the relationship between confidence ratings and correct versus incorrect trials. This led to a significant improvement over the first model, $\chi^2(1) = 45.23$, $p < .001$. Here, we found a significant main effect of Accuracy, ß = -0.91, *SE* = 0.21, $t(358) = -4.30$, $p < .001$, but not Competence, ß < 0.01, *SE* < 0.01, $t(502) = -0.69$, $p = .490$. However, there was a significant Competence x Accuracy interaction, β = 0.017, *SE* < 0.01, $t(397) = 6.85$, $p < .001$. This interaction is illustrated in Figure 6 and shows that the confidence ratings of poor performers failed to discriminate between correct and incorrect responses. In contrast, high performers were more confident in their correct responses. The interaction plotted in Figure 6 directly reflects the central claim of competence-based insight, and the "dual burden" identified by Kruger and Dunning (1999).

Experiment 2

The results of Experiment 1 provided evidence that participants across the sample showed insight into their performance relative to others, and were able to update their estimates post-task in an insightful way. At the trial level, insight (greater confidence in correct versus incorrect responses) was based on competence: better performers showed greater insight. Here, we sought to replicate these results using the same experimental design and using a new, more difficult test of unfamiliar face matching. To this end, we took a confirmatory approach by preregistering this experiment, including its hypotheses, exclusion criteria, and analysis plan (see https://aspredicted.org/blind.php?x=ju5fn9).

Method

Experiment 2 was a preregistered version of Experiment 1 with the only difference being a change from the GFMT (Burton et al., 2010) to the more difficult Kent Face Matching Test (KFMT; Fysh & Bindemann, 2018).

Participants

A representative sample of 463 volunteers (276 women; age M = 30.8 years, SD = 15.7 years; 91% self-reported ethnicity as White) gave informed, onscreen consent before participating in the experiment and were provided with an onscreen debriefing upon completion. Participants were recruited by word of mouth (e.g., through asking friends and family, and sharing the experiment's weblink on social media). There was no overlap between this sample and those who participated in Experiment 1.

The data from an additional four participants were excluded because their scores (42.5%-47.5%) on the KFMT were less than 50%, which represented chance-level performance.

The key finding in Experiment 1 was the interaction between participants' competence on the face matching test and their accuracy (correct/incorrect) when estimating their trial-level confidence ratings. The analysis was carried out using a linear mixed-effects model, and we focussed on this interaction for our power analysis here. Using 'powerSim' (simR package – Green & MacLeod, 2016), we ran simulations with the data from Experiment 1, taking 100 random samples of various numbers of participants and estimating the power based on 20 simulations each. This analysis showed that with 225 participants, we had an average power of 0.96 to detect this interaction. Therefore, we set this as the lower limit for the current experiment, although recruitment continued until the end of a predetermined three-week period.

Stimuli

We used the short version of the KFMT (Fysh & Bindemann, 2018) in order to assess face matching performance. The task comprised 40 pairs of adults (20 male, 20 female), where half the pairs were match trials (different images of the same person – a student ID photograph and a high-resolution portrait) and half were mismatch trials (different people with a similar appearance). All images were in colour and cropped to display only the head and shoulders.

Procedure

The procedure for this experiment was identical to that used in Experiment 1, although here, participants completed the short version of the KFMT as a measure of face matching ability.

In Experiment 1, we tested whether providing participants with examples prior to the task might influence their estimates. Given that we found no evidence that it did, we considered removing this manipulation from Experiment 2. However, the KFMT is a harder test than the GFMT used in Experiment 1 and so there may be scope for the benefit of examples here. As such, we elected to keep the manipulation. Therefore, as in Experiment 1, approximately half of the participants (allocated randomly) were shown an example 'match' trial and 'mismatch' trial (shown in Figure 7), labelled as such, in order to provide additional information about the test. No examples were shown to the remaining participants. Given that the short version of the KFMT included all 20 mismatch trials used in the long version, an example pairing was selected from the Kent University Face Database where neither identity appeared in the test used here. For the example match trial, an item was selected from the long version where the reported difficulty was close to with the mean difficulty of the items in the short version of the test. Again, this identity did not appear in the short version of the test presented to participants.

Results

As noted earlier, we preregistered our analyses and simply replicated those featured in Experiment 1. Overall performance on the KFMT (M = 70.31%, SD = 8.77%, range = 50.0%-92.5%) was comparable with levels found in previous work (66% – Fysh & Bindemann, 2018). Average test scores here were lower than for the GFMT used in Experiment 1 (88%).

As with Experiment 1, although half of the participants were provided with example items before completing the test, we found no evidence that this manipulation affected responses (see the supplementary materials). We therefore collapsed our data across these two subsamples in the analyses that follow.

Traditional Dunning-Kruger analysis using quartiles

As in Experiment 1, we conducted an analysis as in Kruger and Dunning (1999) and Zhou and Jenkins (2020), for the benefit of those wishing to compare our results with other datasets. But to be clear, our theoretical stance is that this analysis is not in itself informative about competence-based insight. For the sake of brevity, we report these analyses in full in the supplementary materials, and illustrate the results in Figure 8. Again, as can be seen in the supplementary materials (Figure S2), the same pattern can be produced by simply shuffling the participants' estimates of their performance (Nuhfer et al., 2016, 2017).

Here, in contrast with Experiment 1, we found that low performers appeared to overestimate while high performers underestimate. This effect being clearly present in Experiment 2, and clearly absent in the conceptually identical Experiment 1, provides further evidence for the effects of perceived task difficulty in generating the intercept of the function for estimated by actual performance (Burson et al., 2006). The KFMT happened to be the right level of difficulty to produce a spread of actual scores that crossed over the line of estimated scores. Our view is that the intercept of the estimated by actual performance function will vary according to overall perceptions of task difficulty. There is little point therefore in trying to identify a performance quartile that is routinely expected to over- or under-perform relative to their estimated performance.

Estimating own performance in relation to others

Participants estimated their own performance, and also how well they thought others would perform on the test, both before and after completing the KFMT. For each participant, we calculated the difference between the participant's estimate of their own performance and that of other people, separately for responses given before and after completing the test

('own_estimate – others_estimate'). As such, positive values represented those who thought they had performed 'better than average'. We then investigated the association between this variable and participants' actual performance on the test.

As we found in Experiment 1, for estimates given before the test, there was a trivial, nonsignificant association with actual performance, $r(461) = .08$, $p = .106$. Unlike Experiment 1, the correlation of actual performance with relative estimates provided after completing the test was also trivial and nonsignificant, $r(461) = .06$, $p = .166$. We are unable to assess the basis of this difference between experiments. The main difference between the two experiments was the difficulty of the face matching test employed, and so it is possible that more difficult tasks may hinder relative performance estimates. In any case, we will not be furthering considering relative performance estimates.

Insight into performance after the test

Next, we investigated participants' updating of their estimated abilities on the KFMT by focussing on how their estimated scores differed from their actual scores. For each participant, we calculated the difference between their estimate of their own score, given before completing the test, and their actual test score (Estimation Error = actual_score – own_estimate_before). We also calculated the difference between their estimates of their own score given before and after the test (Estimation Updating = own_estimate_after – own_estimate_before). Replicating Experiment 1, we found a significant association between these two computed measures, $r(461) = .38$, $p < .001$ (see Figure 9). This association also remained when we considered each quartile separately (lowest: $r = .41$; second: $r = .50$; third: $r = .46$; highest: $r = .44$). As such, participants across our sample demonstrated insight in this task by updating their estimated performance to more accurately reflect their actual performance.

Confidence and competence-based insight

As in Experiment 1, our most important findings relate to the trial-level data analysed using linear mixed-effects models, predicting the confidence rating given by the participant to that trial, given their overall score on the test (Competence) and whether the response was correct or incorrect (Accuracy).

First, we fitted a model in which the interaction between Accuracy and Competence was not included. We found a significant main effect of Accuracy, ß = 0.24, *SE* = 0.02, $t(549) = 12.94$, $p < .001$, such that correct responses were given higher confidence ratings. We also found a nonsignificant effect of Competence, ß = 0.01, *SE* < 0.01, $t(462) = 1.91$, $p = .057$.

Next, we included the Competence x Accuracy interaction to allow for the possibility that participants' abilities influenced the relationship between confidence ratings and correct versus incorrect trials. This led to a significant improvement over the first model, $\chi^2(1) = 14.52$, $p < .001$. Here, we found a significant main effect of Accuracy, ß = -0.31, *SE* = 0.15, $t(486) = -2.14$, $p = .033$, but not Competence, ß < 0.01, *SE* < 0.01, $t(480) = -0.11$, $p = .912$. However, there was a significant Competence x Accuracy interaction, β = 0.008, *SE* < 0.01, $t(516) = 3.83$, $p < .001$. This interaction is illustrated in Figure 10 and shows that, as in Experiment 1, poor performers failed to discriminate between correct and incorrect responses with respect to confidence. In contrast, high performers were more confident in their correct responses. We take this as convincing evidence of competence-based insight.

General Discussion

The main goal of our experiments was to explore the "dual burden" identified by Kruger and Dunning (1999) – that insight into one's performance is based on one's competence – within a domain of social cognition. We wished to take on the arguments made in the past (Burson

et al., 2006; Krueger & Mueller, 2002; Nuhfer et al., 2016, 2017), and move beyond the traditional, quartiles-based analyses used in this field (e.g., Kruger & Dunning, 1999; Zhou & Jenkins, 2020). We therefore considered different ways in which we might measure participants' insights into their own performance in a face matching task. Specifically, we asked participants to provide estimates of their own and others' scores, both before and after completing the test. In addition, we collected a rating of confidence on each trial, alongside their same/different response.

We conducted two similar experiments which varied only in the difficulty of the face matching task. The analyses of Experiment 1 (using the easier GFMT) were exploratory in nature, while Experiment 2 (using the more difficult KFMT) was a preregistered replication. The findings from both experiments were almost identical. We found evidence of insight into social cognition. This insight was both independent of, and dependent upon, competence. Insight independent of competence was observed when participants across both experiments showed accurate updating of their performance estimates. That is, their estimates of how well they *did*, made after completing the test, were closer to their actual performance than their estimates of how well they *would do*, made prior to taking the test, even in the absence of feedback. Participants therefore demonstrated an awareness of their competence during the test, which in turn, allowed refinement of their performance estimates post-completion.

However, our analyses of confidence ratings revealed a degree of insight into performance which *was* dependent upon competence. Those with low competence were equally confident in their correct and incorrect responses. In contrast, high competence participants were more confident in their correct in comparison with their incorrect responses. Previous research has shown that participants, on average, were more confident on trials in which they responded correctly in face matching, recognition, and searching tasks (Bruce et al., 1999; Davis et al., 2018; Grabman & Dodson, 2020; Hopkins & Lyle, 2020; Kramer et

al., 2020; Stephens et al., 2017). However, we have demonstrated here that this was not the case for all participants. Instead, competence on the task determined whether confidence was higher for correct responses or not. This result extends previous work using a test of mathematical knowledge (Händel & Dresel, 2018), where the lowest performers were more confident for incorrect responses, while the highest performers were more confident when their responses were correct. In the current work, the lowest performers did not demonstrate misplaced confidence in their incorrect responses, raising the speculation that even the confidence ratings indicate some awareness in the low-competence individuals. Recent evidence, although within the domains of logical reasoning and grammar, has also supported the notion that low performers are less able to estimate whether they are correct or incorrect on a given trial (Jansen et al., 2021), and this may represent the underlying mechanism behind competence-based insight in general. Further investigation of this trial-level confidence is therefore likely to prove fruitful.

Our findings suggest, perhaps not surprisingly, that insight is a multi-faceted phenomenon. We would characterise our lower performers differently from the classic "unskilled but unaware" label of Kruger and Dunning (1999). All participants, high and low performers alike, demonstrated insight into their overall performance and how that corresponded with their estimates before doing the task. At this global level, poor performers could tell they were having difficulties. What the poor performers could not do, in comparison with the best performers, is identify on a trial-by-trial basis whether they were correct or not. Our poor performers were, therefore, aware they were being overextended, but were likely unable to identify when or why.

That less competent performers were not more confident in their correct (in comparison with their incorrect) responses might be explained in a number of ways. For example, these participants may have already been aware that they tended to perform worse on tests and

therefore felt less confident in general across all of their responses (Fritzsche et al., 2018; Händel & Dresel, 2018). However, our results found no main effect of competence, arguing that poor performers were not overall lower in confidence. Alternatively, less competent performers may have felt confident in some responses and less so on others, but this application of confidence was independent of their response accuracies because they had little insight into their performance.

Our experiments show that in a social domain, competence can predict some but not all forms of insight. Our findings demonstrated insight at the trial level that *was* based on competence. We have also identified measures of insight that were *not* competence-based, with participants of all abilities able to update their estimates in line with their performance on the test. By highlighting issues with some current approaches and demonstrating ways that we can move beyond these traditional methods, we have begun to reveal a more realistic story regarding the nature of insight into social cognition.

References

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ, 332*(7549), 1080.

Anson, I. G. (2018). Partisanship, political knowledge, and the Dunning-Kruger effect. *Political Psychology, 39*(5), 1173-1192.

Aqueveque, C. (2018). Ignorant experts and erudite novices: Exploring the Dunning-Kruger effect in wine consumers. *Food Quality and Preference, 65*, 181-184.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*(1), 37-46.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Bobak, A. K., Mileva, V. R., & Hancock, P. J. (2019). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology, 72*(4), 872-881.

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*(4), 339-360.

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology, 90*(1), 60-77.

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

Davis, J. P., Forrest, C., Treml, F., & Jansari, A. (2018). Identification from CCTV: Assessing police super-recogniser ability to spot faces in a crowd and susceptibility to change blindness. *Applied Cognitive Psychology, 32*(3), 337-353.

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia, 44*(4), 576-585.

Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. *Advances in Experimental Social Psychology, 44*, 247-296.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*(3), 83–87.

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.

Fritzsche, E. S., Händel, M., & Kröner, S. (2018). What do second-order judgments tell us about low-performing students' metacognitive awareness? *Metacognition and Learning, 13*(2), 159-177.

Fysh, M. C., & Bindemann, M. (2018). The Kent Face Matching Test. *British Journal of Psychology, 109*(2), 219-231.

Gignac, G. E., & Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence, 80*, 101449.

Grabman, J. H., & Dodson, C. S. (2020). Stark individual differences: Face recognition ability influences the relationship between confidence and accuracy in a recognition test of Game of Thrones actors. *Journal of Applied Research in Memory and Cognition, 9*(2), 254-269.

Gray, K. L., Bird, G., & Cook, R. (2017). Robust associations between the 20-item prosopagnosia index and the Cambridge Face Memory Test in the general population. *Royal Society Open Science, 4*(3), 160923.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*, 493-498.

Händel, M., & Dresel, M. (2018). Confidence in performance judgment accuracy: The unskilled and unaware effect revisited. *Metacognition and Learning, 13*(3), 265-285.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*(6), 223-233.

Hopkins, R. F., & Lyle, K. B. (2020). Image-size disparity reduces difference detection in face matching. *Applied Cognitive Psychology, 34*(1), 39-49.

Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2021). A rational model of the Dunning-Kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour, 5*(6), 756-763.

Ji, L., & Hayward, W. G. (2020). Metacognition of average face perception. *Attention, Perception, & Psychophysics, 83*, 1036-1048.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. William Collins.

Kelly, K. J., & Metcalfe, J. (2011). Metacognition of emotional face recognition. *Emotion, 11*(4), 896-906.

Krajc, M., & Ortmann, A. (2008). Are the unskilled really that unaware? An alternative explanation. *Journal of Economic Psychology, 29*(5), 724-738.

Kramer, R. S. S. (2021). Forgetting faces over a week: Investigating self-reported face recognition ability and personality. *PeerJ, 9*, e11828.

Kramer, R. S. S., Hardy, S. C., & Ritchie, K. L. (2020). Searching for faces in crowd chokepoint videos. *Applied Cognitive Psychology, 34*(2), 343-356.

Kramer, R. S. S., Mohamed, S., & Hardy, S. C. (2019). Unfamiliar face matching with driving licence and passport photographs. *Perception, 48*(2), 175-184.

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

Krueger, I. J., & Mueller, A. R. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology, 82*, 180-188.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

Kuhn, D. (2000). Theory of mind, metacognition, and reasoning: A life-span perspective. In P. Mitchell & K. J. Riggs (Eds.), *Children's reasoning and the mind* (pp. 301-326). Psychology Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1-26.

Livingston, L. A., & Shah, P. (2018). People with and without prosopagnosia have insight into their face recognition ability. *Quarterly Journal of Experimental Psychology, 71*(5), 1260-1262.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*(3), 280–296.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19-40.

Matsuyoshi, D., & Watanabe, K. (2021). People have modest, not good, insight into their face recognition ability: A comparison between self-report questionnaires. *Psychological Research, 85*, 1713-1723.

McClelland, G. H., Lynch, J. G., Jr., Irwin, J. R., Spiller, S. A., & Fitzsimons, G. J. (2015). Median splits, Type II errors, and false–positive consumer psychology: Don't fight the power. *Journal of Consumer Psychology, 25*(4), 679-689.

Meeran, S., Goodwin, P., & Yalabik, B. (2016). A parsimonious explanation of observed biases when forecasting one's own performance. *International Journal of Forecasting, 32*(1), 112-120.

Nuhfer, E., Cogan, C., Fleisher, S., Gaze, E., & Wirth, K. (2016). Random number simulations reveal how random noise affects the measurements and graphical portrayals of self-assessed competency. *Numeracy, 9*(1), 4.

Nuhfer, E., Fleisher, S., Cogan, C., Wirth, K., & Gaze, E. (2017). How random noise and a graphical convention subverted behavioral scientists' explanations of self-assessment data: Numeracy underlies better alternatives. *Numeracy, 10*(1), 4.

Pavel, S. R., Robertson, M. F., & Harrison, B. T. (2012). The Dunning-Kruger effect and SIUC University's aviation students. *Journal of Aviation Technology and Engineering, 2*(1), 125-129.

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2017). Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review, 24*(6), 1774-1784.

Richardson, H., & Saxe, R. (2020). Development of predictive responses in theory of mind brain regions. *Developmental Science, 23*(1), e12863.

Ross, L., Greene, D., & House, P. (1977). The false consensus effect: An egocentric bias in social perception and attributional processes. *Journal of Experimental Social Psychology, 13*, 279-301.

Shah, P., Sowden, S., Gaule, A., Catmur, C., & Bird, G. (2015). The 20 item prosopagnosia index (PI20): Relationship with the Glasgow face-matching test. *Royal Society Open Science, 2*(11), 150305.

Stephens, R. G., Semmler, C., & Sauer, J. D. (2017). The effect of the proportion of mismatching trials and task orientation on the confidence–accuracy relationship in

unfamiliar face matching. *Journal of Experimental Psychology: Applied, 23*(3), 336-353.

Ventura, P., Livingston, L. A., & Shah, P. (2018). Adults have moderate-to-good insight into their face recognition ability: Further validation of the 20-item Prosopagnosia Index in a Portuguese sample. *Quarterly Journal of Experimental Psychology, 71*(12), 2677-2679.

Zhou, X., & Jenkins, R. (2020). Dunning-Kruger effects in face perception. *Cognition, 203*, 104345.

Figures

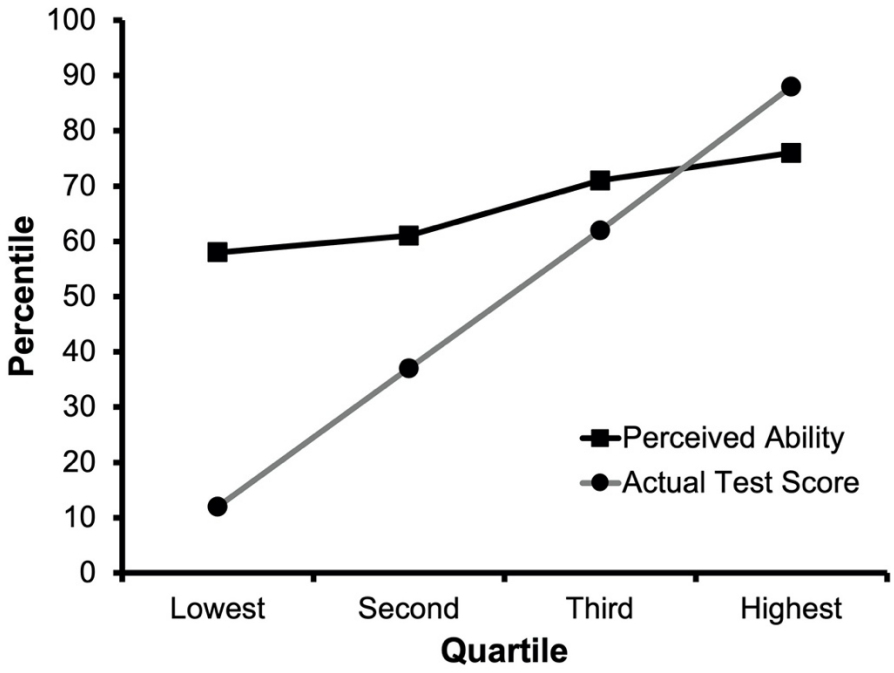

Figure 1. The classic Dunning-Kruger Effect, illustrating the results from Study 1 of Kruger and Dunning (1999). Estimated performance is imperfectly correlated with actual performance, such that the lowest performing quartile shows the greatest overestimate of their actual ability.

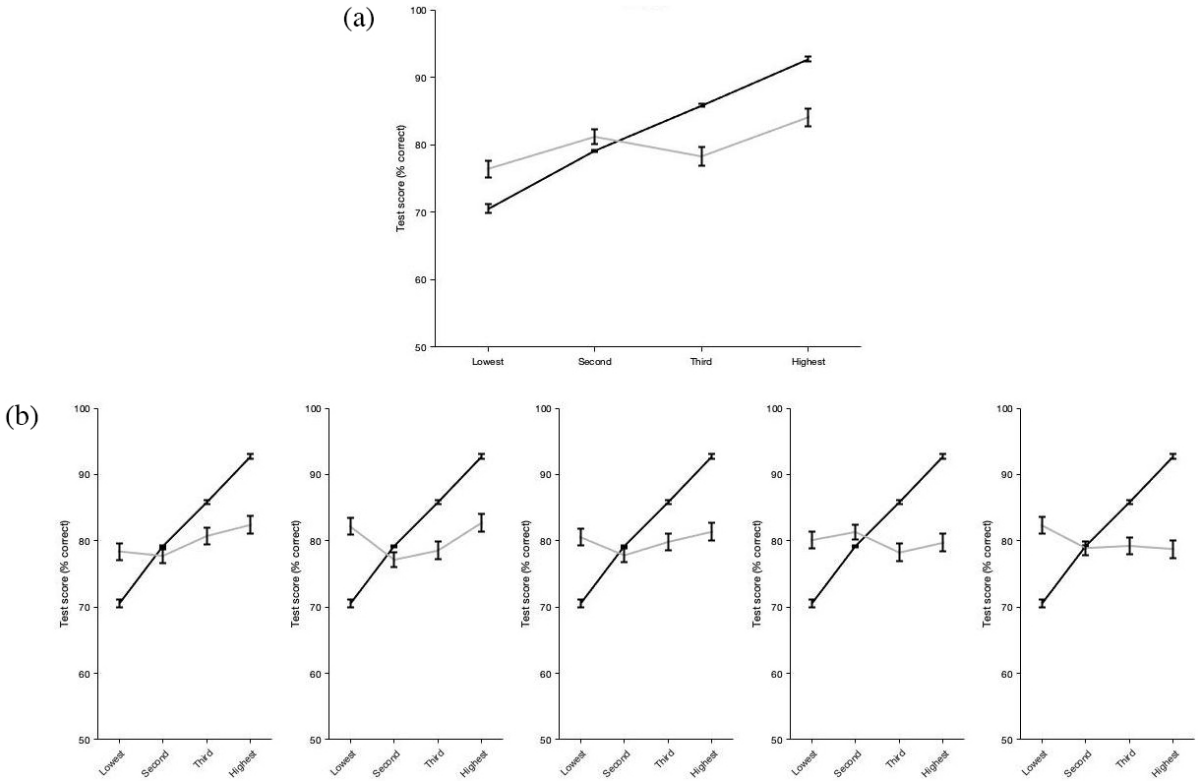

Figure 2. Estimated (grey) and actual (black) performance across quartiles for (a) the original pattern of results in Zhou and Jenkins (2020) and (b) five iterations of shuffled data. Error bars represent standard error.

Figure 3. Example face pairs from the GFMT. A match pair (top row) and a mismatch pair (bottom row).

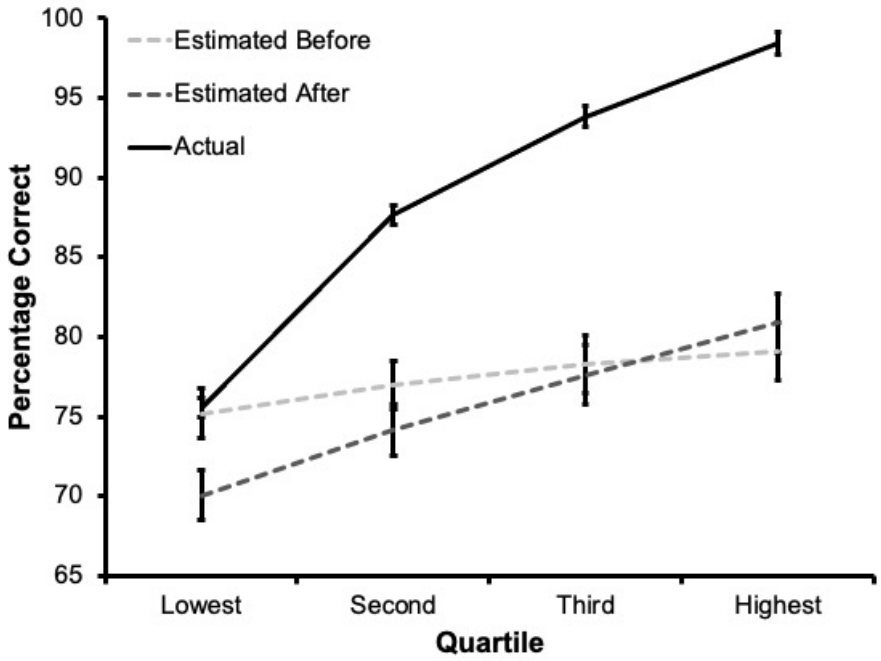

Figure 4. A summary of both estimated and actual GFMT performance across quartiles. Error bars represent 95% confidence intervals.

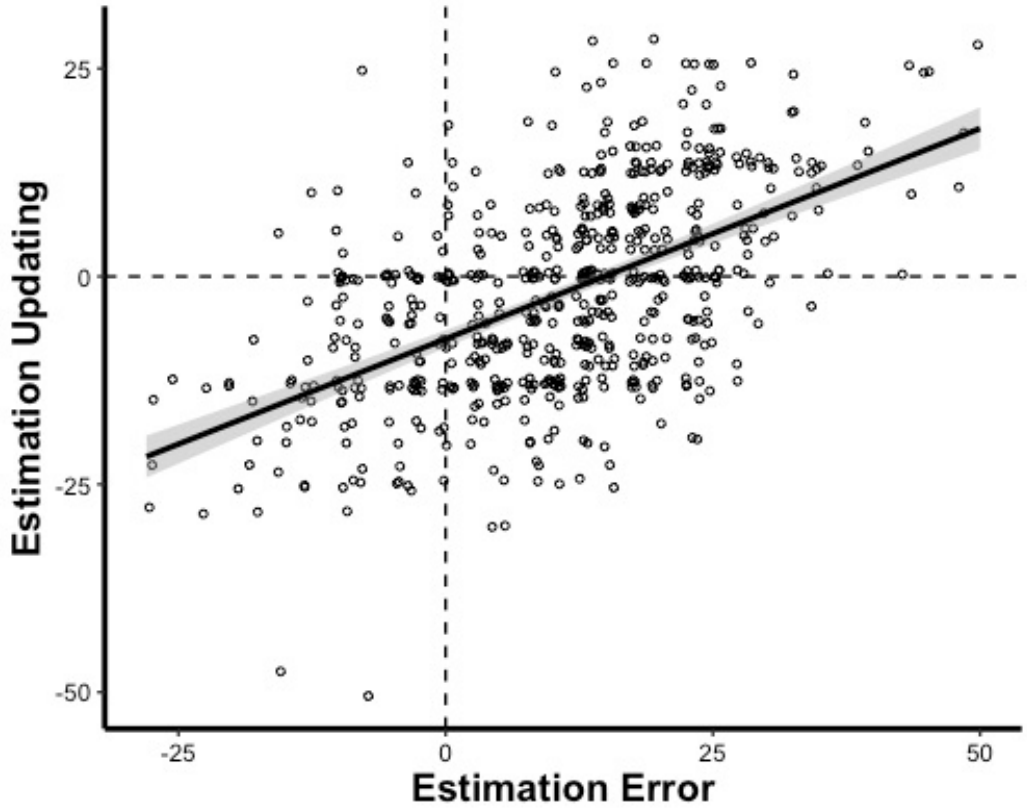

Figure 5. The association between estimation error and updating across our sample. The black line represents the linear model and the error band represents the 95% confidence interval.

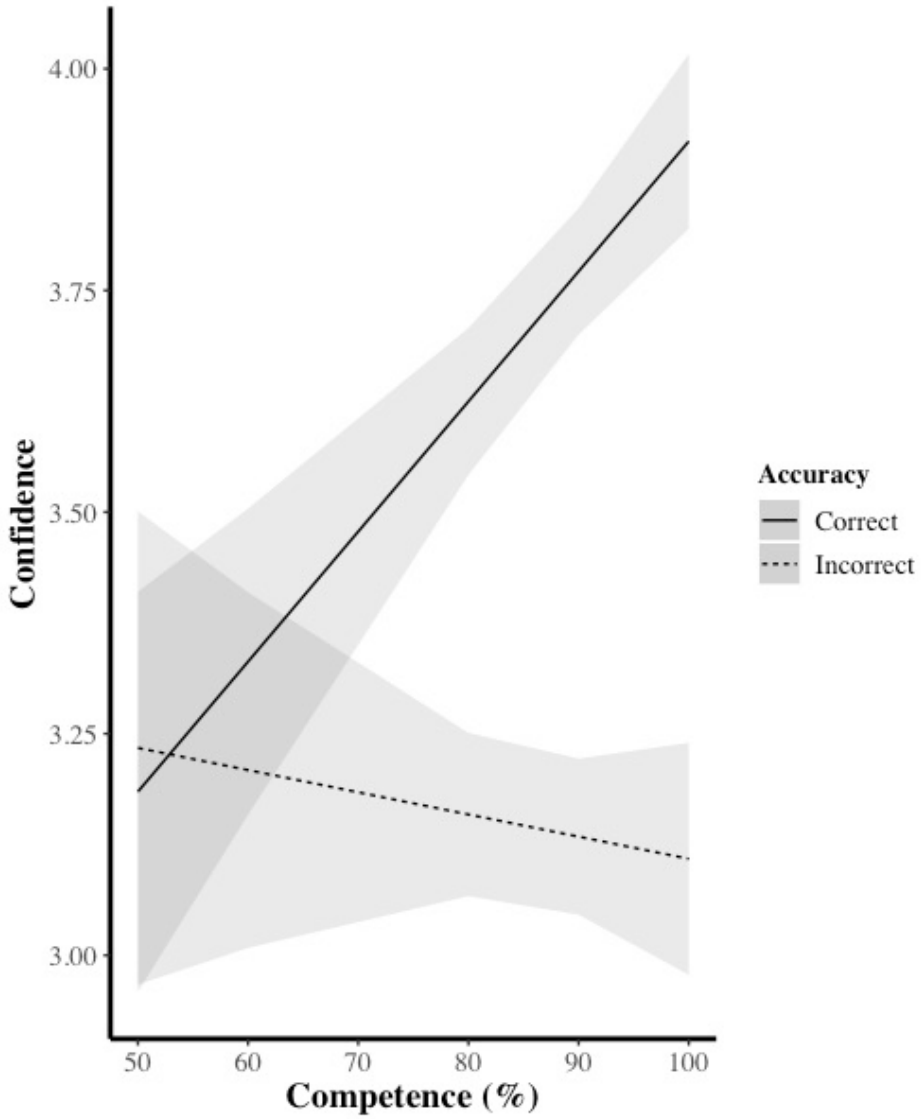

Figure 6. An illustration of confidence as a function of competence in the model, separately for correct and incorrect responses. Error bands represent 95% confidence intervals.

Figure 7. Example face pairs from the KFMT. A match pair (top row) and a mismatch pair (bottom row).

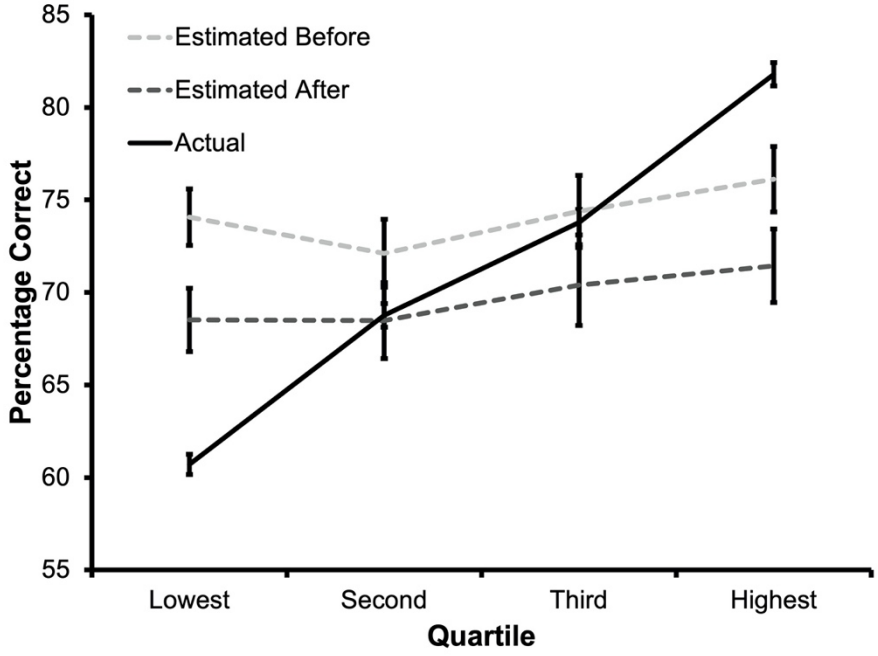

Figure 8. A summary of both estimated and actual KFMT performance across quartiles. Error bars represent 95% confidence intervals.

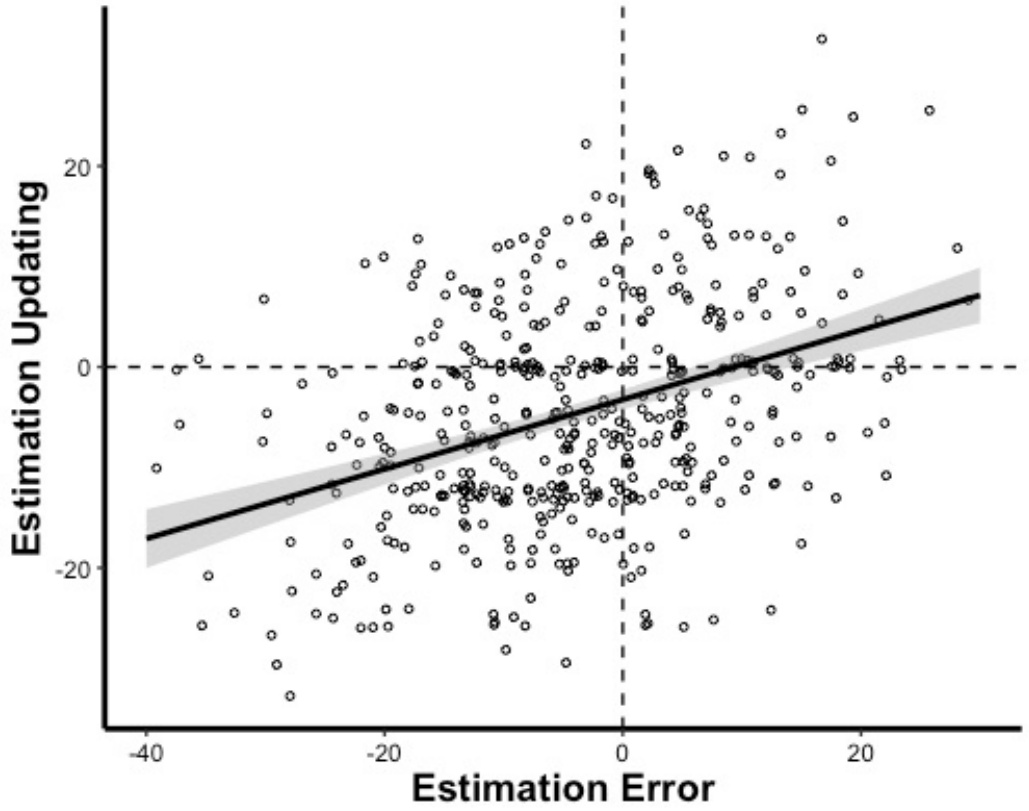

Figure 9. The association between estimation error and updating across our sample. The black line represents the linear model and the error band represents the 95% confidence interval.

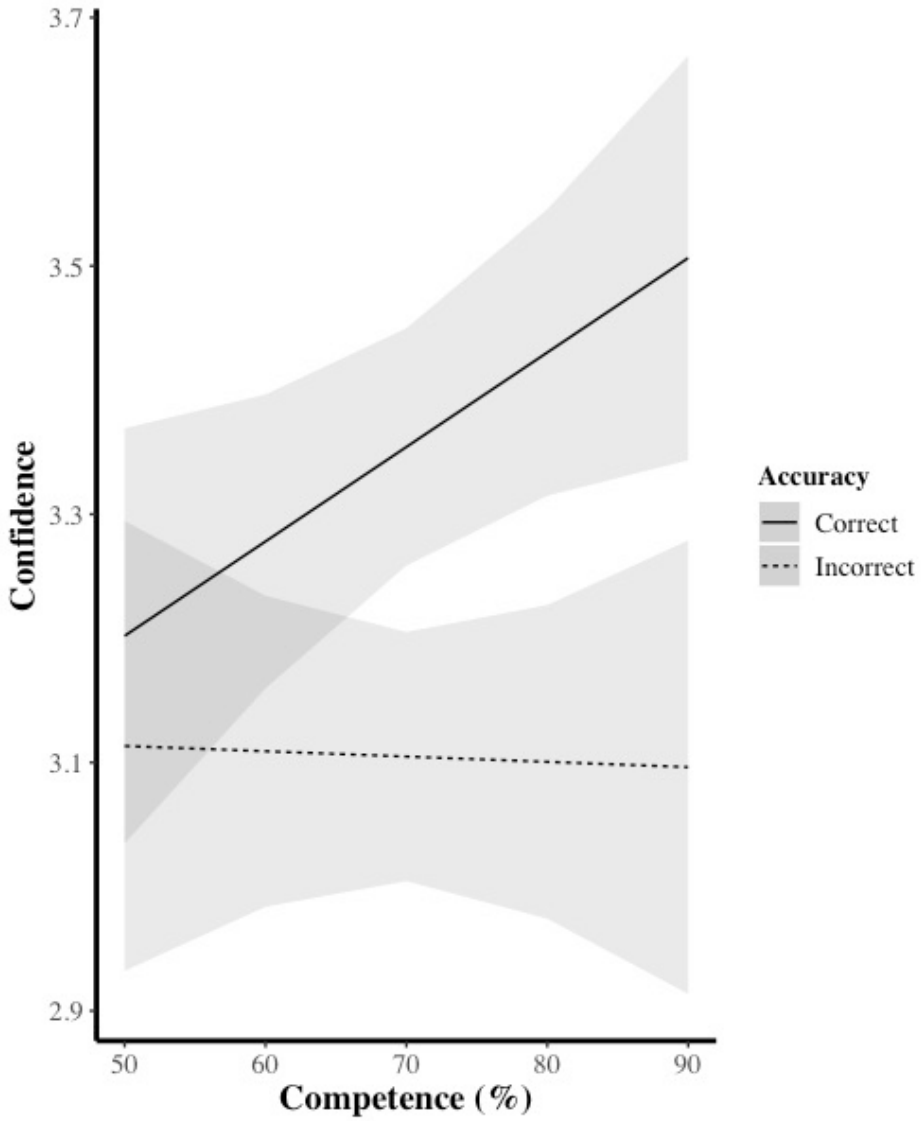

Figure 10. An illustration of confidence as a function of competence in the model, separately for correct and incorrect responses. Error bands represent 95% confidence intervals.