THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

The (mis)use of performance quartiles in metacognition and face perception

A comment on Zhou and Jenkins (2020) and Estudillo and Wong (2021)

Citation for published version:
Kramer, RSS, McIntosh, RD & Nuhfer, EB 2023, 'The (mis)use of performance quartiles in metacognition and face perception: A comment on Zhou and Jenkins (2020) and Estudillo and Wong (2021)', *Psychological Reports*. https://doi.org/10.1177/00332941231181483

Digital Object Identifier (DOI):
10.1177/00332941231181483

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Psychological Reports

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 13. Dec. 2024

The (mis)use of performance quartiles in metacognition and face perception: A comment on Zhou and Jenkins (2020) and Estudillo and Wong (2021)

Robin S. S. Kramer[1], Robert D. McIntosh[2], and Edward B. Nuhfer[3]

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

[2]Human Cognitive Neuroscience, Psychology, University of Edinburgh, Edinburgh, UK

[3]California State Universities (Humboldt Polytech, Arcata CA, USA, retired)

Short title: Performance quartiles in metacognition and face perception

Corresponding author:

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

Abstract

A common measurement convention within the field of metacognition is to divide participants into quartiles based on task performance, and then compare self-estimated and actual scores within these sub-groups. This analysis strategy created the famous Dunning-Kruger effect, which asserts that the poorest performers tend to grossly overestimate their abilities. A study by Zhou and Jenkins (2020) has recently replicated this effect within the domain of face matching. However, it can be shown that the analysis strategy induces numerical artefacts prone to misinterpretation, and that randomly generated data lead to the same pattern of results. Estudillo and Wong (2021) used a different quartiles-based approach to argue that only the lowest and highest performers on a task of face recognition showed some insight into their performance. Again, a numerical artefact can explain their result, with the restricted range of the second and third quartiles causing reduced observed correlations between actual and self-estimated abilities. These studies highlight the need for methodological caution when exploring metacognitive questions, and we outline some avenues that may aid future investigation.

Keywords

metacognition, quartiles, Dunning-Kruger effect, face perception, face matching, regression to the mean

Introduction

This commentary draws attention to two problematic analysis strategies in recent studies of the metacognition of face perception. The first strategy is prevalent in the wider literature on metacognition, and we should be vigilant not to let it become established within the face perception literature. The second is relatively novel, and so it may be useful to raise awareness of its shortcomings here. Both strategies involve studying the relationship between paired measures by dividing a sample into quartiles on one of those measures, and both strategies induce numerical artefacts that may be mistaken by the unwary for meaningful patterns of behaviour. Our commentary focusses on articles by Zhou and Jenkins (2020) and Estudillo and Wong (2021), which illustrate these problematic analysis strategies, as applied to face perception.

The Dunning-Kruger effect and regression to the mean

Metacognition refers to the capacity to reflect on and assess one's own cognitive processes. An obvious way to study this capacity is to elicit self-estimates of performance from participants on a particular task and compare these estimates to measures of their actual performance. We might further wish to know whether people

with different levels of ability on a task differ in terms of their metacognitive insight. Kruger and Dunning (1999) developed a method for making this comparison that has since become common. Unfortunately, the method is dangerous because it is liable to produce numerical artefacts that can masquerade as metacognitive differences between the best and the worst performers.

Kruger and Dunning obtained actual and self-estimated measures of performance by asking participants to complete a cognitive task (for instance, logical reasoning), and also to estimate their level of performance, either as an absolute score or as a percentile relative to others. Participants were then ranked by actual score and divided into four quartiles of ability from worst (lowest quartile) to best (highest quartile). The data were visualised using the graphical convention shown in Figure 1, which illustrates how self-estimated performance deviates from actual performance across the range of abilities. When this is done, the worst performers invariably overestimate themselves to a far greater extent than the best performers, who are generally more accurate in their self-assessments and may even underestimate themselves. Figure 1a reproduces the first implementation of this approach (Study 1; Kruger & Dunning, 1999).

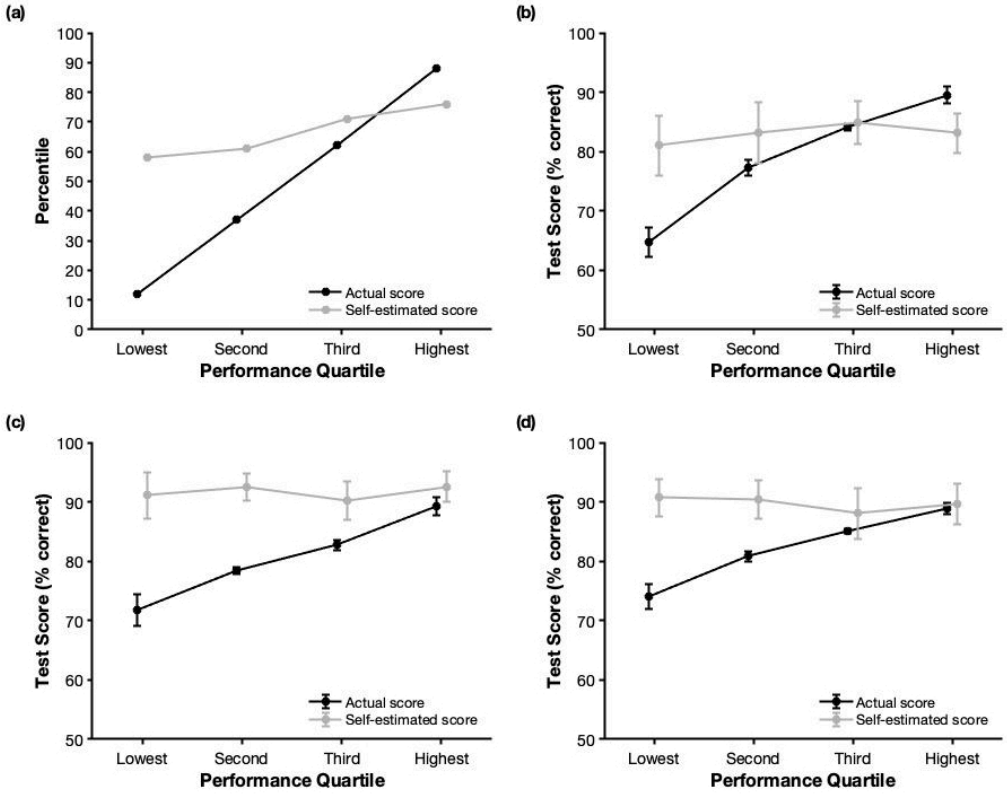

Figure 1. The classic DKE graphical convention, illustrating (a) the results from Study 1 of Kruger and Dunning (1999); and the results from Experiment 2 of Zhou and Jenkins (2020) for (b) face identity, (c) gaze direction, and (d) emotion expression. Error bars represent 95% confidence intervals (not available for panel a).

This general pattern of data communicates the Dunning-Kruger effect (DKE; Dunning, 2011), now replicated numerous times across diverse cognitive domains and tasks (some recent examples include: Anson, 2018; Aqueveque, 2018; Greitemeyer, 2020; Lyons et al., 2021; Pennycook et al., 2017). Likewise, Zhou and Jenkins (2020)

have reported that the same pattern holds for multiple aspects of face perception, including the matching of identity, gaze direction, and emotional expression (Figures 1b–1d). The lowest quartile of performers significantly overestimated themselves for each of these tasks, with the authors concluding that a lack of face processing ability was accompanied by a lack of insight into one's ability.

However, this pattern always arises from a DKE analysis, even if the data are just bounded sets of random numbers (Magnus & Peresetsky, 2022; for further discussion, see Kramer et al., 2022). This is illustrated in Figure 2a, which shows the same method of analysis as applied to random uncorrelated variables X and Y (to stand in for actual and self-estimated performance respectively). If we focus on those participants who scored in the lowest quartile on X, we will find that their score on Y is not nearly as extreme; and the same will be true for participants in the highest quartile (but in the opposite direction). This is a version of the familiar DKE in that lowest quartile participants overestimate themselves more than the highest quartile participants (who, in fact, underestimate themselves), but we would be wrong to interpret this in terms of metacognitive differences. Instead, it is the result of regression to the mean, an inevitable numerical consequence of the fact that we selected subgroups of participants for being extreme on one variable (X), and then evaluated their score on a second variable (Y) with respect to the first. The DKE analytical strategy is a recipe for

regression to the mean, which is why the effect replicates so reliably across diverse domains.

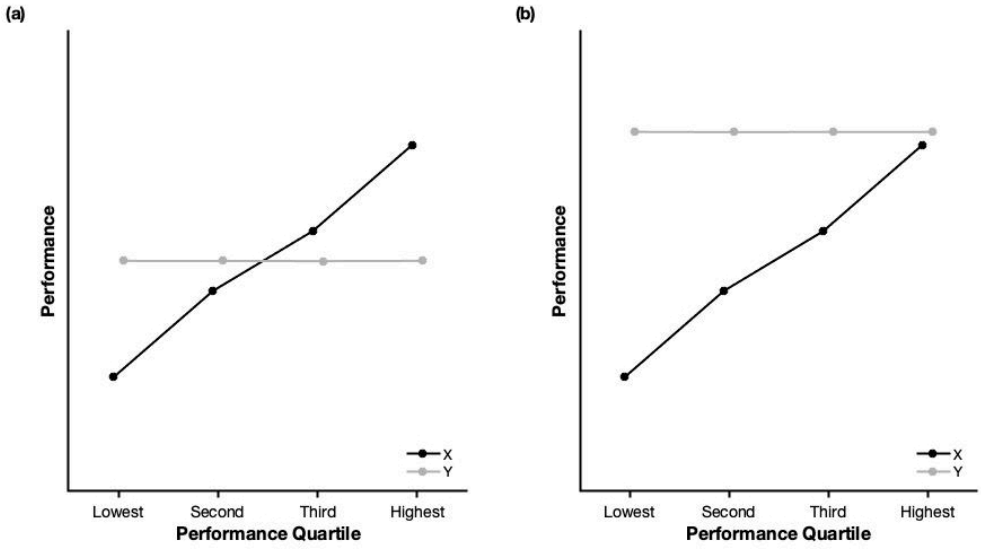

Figure 2. Randomly generated datasets of paired data with a correlation of zero. (a) Both X and Y are generated from distributions with the same mean and standard deviation. (b) Y is generated from a distribution with the same standard deviation as X, but a higher mean. Quartile means come from averaging over 10,000 iterations. Error bars representing 95% confidence intervals are negligible in length.

Unlike in Figure 2a, the classic DKE combines overestimation amongst the lowest-quartile performers with more accurate self-estimation in the highest quartile. However, this is easily modelled by setting the mean value of Y (self-estimation) to be

higher than the mean value of X (actual performance), as has been done in Figure 2b. In general, if participants think that the task is fairly easy, then their mean self-estimates will be quite high and overestimation at the lowest end will be exaggerated, whilst the self-estimates of the highest performers will look more accurate (Burson et al., 2006). Zhou and Jenkins (2020) observed very high mean self-estimates for matching in relation to gaze direction (Figure 1c) and emotional expression (Figure 1d), suggesting that participants generally thought these were quite easy tasks. However, the self-estimate is not necessarily a good measure of metacognition because such methods cannot distinguish between metacognitive bias, a general tendency to give high or low estimates, and metacognitive sensitivity, an ability to discriminate whether one is performing well or poorly on a given trial (Fleming & Lau, 2014).

At the participant level, metacognitive sensitivity cannot be measured by global or aggregate self-estimates. For this, we need psychophysical analyses of how well self-estimates track performance across trials (see Fleming & Lau, 2014). Alternatively, we can examine whether self-estimates show any sensitivity to actual performance across participants by studying the correlation of self-estimated and actual performance. The correlations that Zhou and Jenkins (2020) found for unfamiliar face perception tasks were low (-0.06 to 0.27), as was also the case in Kruger and Dunning's (1999) original studies (of humour, grammar, and reasoning). The general lack of a relationship is indicated by the near-horizontal grey lines seen in Figure 1, which differ little from

those for the random, uncorrelated variables in Figure 2. This lack of correlation implies that participants have little or no metacognitive insight into their face processing abilities and/or that aggregate or one-shot estimates are not valid measures of metacognitive insight.

It is important to emphasise that even if more substantial correlations were observed between self-estimated and actual performance (resulting in grey lines with positive slopes), the problem of regression to the mean would not be resolved, although the severity of its effects might be reduced. Regression to the mean will always apply to any two variables unless they are perfectly correlated. The best way to deal with regression to the mean is to be aware of it, and to avoid analyses that could make our results prone to its influence (Campbell & Kenny, 2002). Kruger and Dunning's (1999) method of analysis by quartiles is especially prone to regression, which has been highlighted repeatedly in the scientific literature (e.g., Burson et al., 2006; Gignac & Zajenkowski, 2020; McIntosh et al., 2019; Nuhfer et al., 2016, 2017) and recently in an article aimed at a wider readership (McIntosh & Della Sala, 2022). Researchers interested in metacognition in relation to face perception should be aware of this statistical artefact and avoid making causal explanations from it.

Within-quartile correlations and range effects

As noted, one potential strategy to assess metacognitive sensitivity is to study the correlation between self-estimates and actual performance. Provided that the self-estimate is a valid measure of metacognitive insight, then this correlation can be informative about metacognitive sensitivity at the group level. This approach was taken recently by Estudillo and Wong (2021), who were interested in the relationship between face recognition ability, assessed by the Chinese version of the Cambridge Face Memory Test (CFMT-Chinese; McKone et al., 2012), and insight into one's own face recognition difficulties, assessed by the 20-item prosopagnosia index (PI-20; Shah, Gaule, et al., 2015; Shah, Sowden, et al., 2015). Across a sample of 255 Chinese ethnicity students, they found a moderate overall association between test scores and self-estimated abilities (r = -0.35).[1]

In an attempt to investigate whether the level of insight varied across the range of performance, Estudillo and Wong then subdivided their sample into performance quartiles on the ability test (CFMT) and re-assessed the correlation with PI-20 scores per quartile. They found that the group-level relationship remained statistically significant for the lowest and highest quartiles only. They also replicated this result in a secondary dataset reanalysed from Gray et al. (2017; N = 425). Given this consistent pattern across quartiles, the authors concluded that only people at the lowest or highest ends of actual ability have metacognitive insight into their level of performance.

[1] Because high CFMT scores indicate better face recognition abilities, and higher PI-20 scores represent more subjective face processing difficulties, the expected correlation is negative.

However, it is once again the numerical approach that explains these apparent subgroup differences, rather than any true differences of metacognition. To appreciate this, it is critical to consider the effect of range restriction on the strength of correlation (see Goodwin & Leech, 2006; also noted by Burson et al., 2006). By dividing the face recognition test scores (assumed to be normally distributed within the population) into quartiles, we should expect that the scores within the second and third quartiles (the middle 50% of the data) are more restricted in range than those within the lowest and highest quartiles (see Figure 3).

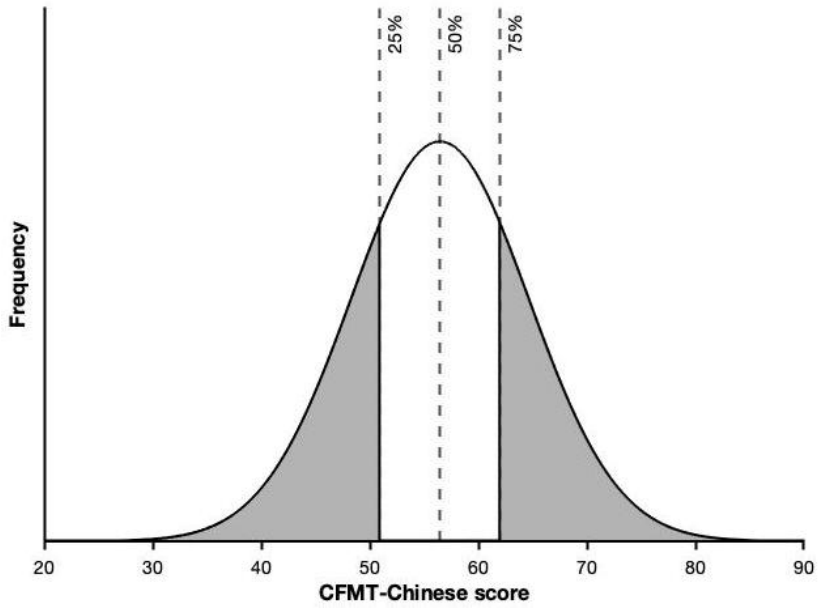

Figure 3. Quartiles and their score ranges. For the distribution of CFMT-Chinese test scores (M = 56.5, SD = 8.2; Estudillo & Wong, 2021), as for any normal distribution,

the second and third quartiles (white) will have a more restricted range of test scores (spanning 0.67 standard deviations either side of the mean) than the lowest and highest quartiles (extending out to more than 2 standard deviations; grey).

For any bivariate population with a non-zero underlying relationship, restricting the range of a subsample will result in a lower estimated correlation. Conversely, because the spread of data in the lowest and highest quartiles is greater than in the middle quartiles, we should expect to see larger correlations within those quartiles. As Figure 4 illustrates, randomly generated bivariate data with a population-level correlation of -0.35 reproduce exactly this pattern of enhanced correlations in the lowest and highest quartiles (though note that all within-quartile correlations are weaker than the true population correlation, again due to range restriction).[2]

[2] Well-established methods exist for addressing relationships between variables bounded by restricted ranges. For example, Thorndike's Case 2 adjustment can be used to predict the association across the whole sample when faced with range-restricted data (see Sackett & Yang, 2000).

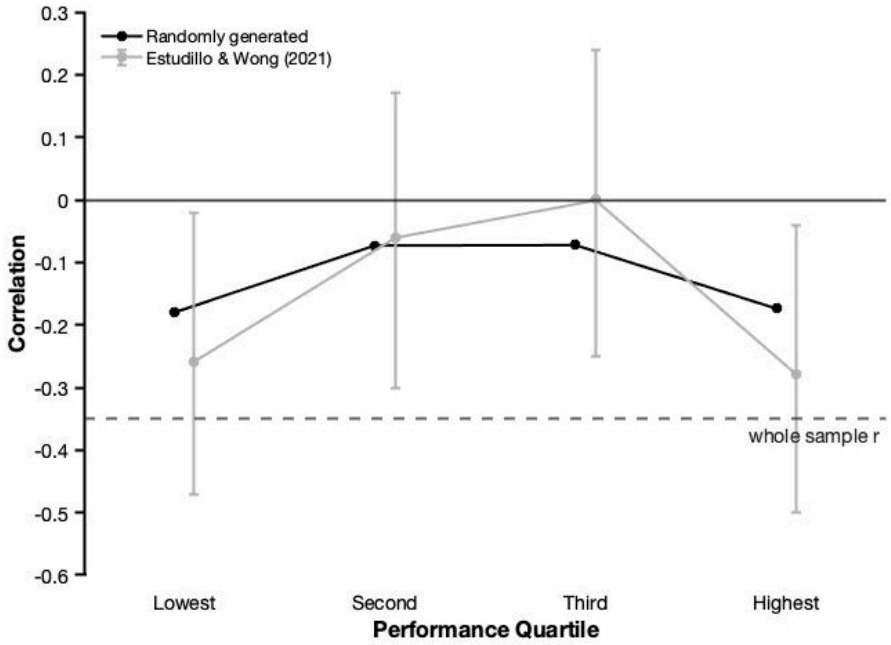

Figure 4. The pattern of quartile correlations resulting from randomly generated data following the parameters of Estudillo and Wong (2021). In each iteration, the actual scores and self-estimated abilities were constrained to show a correlation of -0.35 (dashed line) and came from normal distributions with means and standard deviations matching the original data. Quartile correlations are averaged over 10,000 iterations. The original data from Estudillo and Wong appear in grey for comparison. Error bars represent 95% confidence intervals and are negligible in length for the randomly generated data.

This range restriction artefact is quite distinct from regression to the mean discussed above, although both may be magnified by subdividing participants into

quartiles. More generally, such arbitrary divisions of continuous variables, such as the common clinical practice of dichotomising, can mask group level patterns and reduce power to detect true relationships in the dataset (Altman & Royston, 2006; MacCallum et al., 2002; McClelland et al., 2015).

Moving beyond quartiles

After accepting that analytical strategies based on the arbitrary sorting and division of continuous data into performance quartiles are potentially problematic, how might we proceed when investigating metacognitive insight within face perception? To explore the relationship between actual scores and self-estimates of ability, one must first establish the reliability of the data produced from each of the two measures. If either measure fails to produce data with acceptable levels of reliability then advancing to the stage of paired comparisons is premature (Nuhfer et al., 2016). In paired measures, the two instruments/measures must also be aligned, meaning that participants self-assess their competence on the same challenge in which they demonstrate competence.

Within the domain of face perception, researchers have arguably yet to establish a measure of self-reported ability that is sufficiently aligned with actual performance measures (Bobak et al., 2019; Matsuyoshi & Watanabe, 2021). For example, the correlation between performance and derived self-estimates featured in Zhou and

Jenkins (2020) was 0.27 or below, while the strength of (negative) correlation between CFMT and PI-20 scores in Estudillo and Wong (2021) was 0.35. Such misalignment suggests either that metacognitive insight for face perception is very poor in general or that the measures themselves are insufficiently well-developed.

A second approach is to focus on trial-level insight. By collecting test responses along with confidence ratings for those responses, researchers have shown that individuals who performed better on tasks of face matching and recognition were also those demonstrating significantly higher confidence in their correct responses than in their incorrect ones (e.g., Grabman & Dodson, 2022; Kramer, 2023; Kramer et al., 2022). In contrast, poor performers' confidence ratings failed to differentiate between their correct and incorrect responses. These findings suggest that people with poorer face recognition abilities do indeed have poorer metacognitive insight, in the sense that they are less able to distinguish between their successes and failure. This is consistent with one premise of the DKE, but it contradicts the associated idea that poor performers are overconfident in their abilities. Very similar results were also reported for a logical reasoning test, suggesting that these patterns may generalise well across domains (McIntosh et al., 2022).

To conclude, a quartiles-based approach to analysis seems pervasive within the literature on metacognition (particularly in studies of the DKE) and has recently appeared in the domain of face perception. However, flaws and pitfalls associated with

this type of analysis are now well documented, and we strongly urge future researchers to pursue other methods when assessing metacognitive insight for face perception.

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

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.

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.

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.

Campbell, D. T., & Kenny, D. A. (2002). *A primer on regression artifacts*. Guilford Press.

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

Estudillo, A. J., & Wong, H. K. (2021). Associations between self-reported and objective face recognition abilities are only evident in above- and below-average recognisers. *PeerJ, 9*, e10629.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, 443.

Grabman, J. H., & Dodson, C. S. (2022*). Unskilled, underperforming, or unaware? Testing three accounts of individual differences in metacognitive monitoring sensitivity*. ResearchGate. https://doi.org/10.13140/RG.2.2.21924.35208

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.

Greitemeyer, T. (2020). Unattractive people are unaware of their (un)attractiveness. *Scandinavian Journal of Psychology, 61*(4), 471-483.

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.

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of *r*. *The Journal of Experimental Education, 74*(3), 249-266.

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

Kramer, R. S. S., Gous, G., Mireku, M. O., & Ward, R. (2022). Metacognition during unfamiliar face matching. *British Journal of Psychology, 113*(3), 696-717.

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.

Lyons, B. A., Montgomery, J. M., Guess, A. M., Nyhan, B., & Reifler, J. (2021). Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences, 118*(23), e2019527118.

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.

Magnus, J. R., & Peresetsky, A. A. (2022). A statistical explanation of the Dunning–Kruger effect. *Frontiers in Psychology, 13*, 840180.

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.

McIntosh, R. D., & Della Sala, S. (2022). The persistent irony of the Dunning-Kruger Effect. *The Psychologist, 35*(3), 30-34.

McIntosh, R. D., Fowler, E. A., Lyu, T., & Della Sala, S. (2019). Wise up: Clarifying the role of metacognition in the Dunning-Kruger effect. *Journal of Experimental Psychology: General, 148*(11), 1882-1897.

McIntosh, R. D., Moore, A. B., Liu, Y., & Della Sala, S. (2022). Skill and self-knowledge: Empirical refutation of the dual-burden account of the Dunning–Kruger effect. *Royal Society Open Science, 9*(12), 191727.

McKone, E., Stokes, S., Liu, J., Cohan, S., Fiorentini, C., Pidcock, M., Yovel, G., Broughton, M., & Pelleg, M. (2012). A robust method of measuring other-race

and other-ethnicity effects: The Cambridge Face Memory Test format. *PLoS ONE, 7*(10), e47956.

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.

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.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85*(1), 112-118.

Shah, P., Gaule, A., Sowden, S., Bird, G., & Cook, R. (2015). The 20-item prosopagnosia index (PI20): A self-report instrument for identifying developmental prosopagnosia. *Royal Society Open Science, 2*(6), 140343.

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.

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