CORTEX 186 (2025) 86–98

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research Report

Exploring insight into unfamiliar face recognition ability: The case from developmental prosopagnosia

Jeremy J. Tree * *and Alex L. Jones*

School of Psychology, University of Swansea, Swansea, UK

ARTICLE INFO

Article history:
Received 6 January 2025
Revised 25 March 2025
Accepted 25 March 2025
Action editor: Sascha Frühholz

ABSTRACT

This study aims to explore the relationship between face processing ability and individuals' *insight* into that ability, with a particular focus on those who 'self-refer' as having face recognition difficulties; namely, individuals with developmental prosopagnosia (DP). Specifically, the study examines whether self-referred individuals represent a subpopulation with elevated levels of insight into their face recognition performance compared to the general population. Using Bayesian hierarchical modelling, we compared performance across the 'objective' Cambridge Face Memory Test (CFMT) and the 'subjective' 20-item Prosopagnosia Index (PI20) in self-referred DP individuals (N = 279) and normative populations (N = 1,344)-this statistical approach allows for flexible, probabilistic predictions about performance based on subjective insight and group membership, enabling more nuanced comparisons. Despite hypotheses that self-referring participants might demonstrate superior metacognitive insight, results showed no credible evidence of enhanced alignment between PI20 and CFMT measures in this group compared to normative samples. Overall, these findings underscore the limitations of current diagnostic tools, emphasizing the need for psychometric refinement to address measurement noise and improve the reliability of subjective self-assessments. This work contributes to understanding individual variability in cognitive insight and highlights the challenges of identifying DP based on subjective and objective alignment.

1. Introduction

The ability to recognise faces is a critical cognitive skill, relevant to a large variety of aspects of the human experience. It is hard to imagine how one may do without it in day-to-day life-yet, there is a surprisingly large degree of individual differences in performance on standardised measures of face recognition (such as the Cambridge Face Memory Test-(CFMT)-Duchaine & Nakayama, 2006)-to the extent that individuals at the 'extremes' of this population distribution of performance have been named as distinct groups;

* *Corresponding author*. School of Psychology, Faculty of Medicine, Health and Life Sciences, University of Swansea, Swansea SA2 8PP, UK.
E-mail address: j.tree@swansea.ac.uk (J.J. Tree).
https://doi.org/10.1016/j.cortex.2025.03.009

developmental prosopagnosia (DP) for those doing very poorly (e.g., Bate, Bennetts, Tree, et al., 2019,b; Bennetts et al., 2022), and *super-recognisers (SR)* for those doing very well (e.g., Bobak et al., 2016; Davis et al., 2016). Early work in the context of DP was important for setting the foundations of understanding that individuals with 'extremely' poor performance were by no means as rare as assumed, that their poor ability was likely face-specific (Towler & Tree, 2018) and that within this population the nature of the *functional* cause for their face processing impairment likely varied (i.e., the *nature* of their prosopagnosia is heterogeneous-Jackson et al., 2017, Wilcockson et al., 2020, Burns et al., 2014). As a consequence, discussions shifted to the observation that some kind of *convergence* on tests to 'diagnose' (i.e., select key participants) needed to be adopted (Bate & Tree, 2017; Nørkær, Gobbo, Roald, & Starrfelt, 2024). In the spirit of such *test battery* development, researchers have created tests that largely fall into experimental-based 'objective' behavioural performance measures (e.g., that might measure unfamiliar face recognition memory: such as the CFMT) and 'subjective' self-report measures (linked to the 'day-to-day' face processing experience of individuals: such as the 20 item prosopagnosia index (PI20)-Shah et al., 2015). This makes rational sense since many 'diagnostic' constructs (e.g., sub-types of dementia) involve clinicians gathering *both* during clinical interviews (e.g., Jenkins et al., 2015, Jenkins, Tree, Thornton & Tales, 2019).

However, although it would be ideal to have self-report measures of day-to-day face processing problems alongside other behavioural measures to select potential DP candidates for research, there is a consistent issue with this approach, which is that these two forms of measurement *rarely* correlate well. For example, studies that have reported general population correlations between key 'objective' performance (CFMT) and 'subjective' (PI20) face measures, clearly indicate variable levels of association (see examples in Table below), with r values varying from −.67 (Shah and colleagues who did the original PI20 work), to as low as −.16 (Stantic et al., 2021; Expt 3)-note that the correlation is negative because *low* scores on the CFMT and *high* scores on the PI20 respectively reflect 'poor' ability across both measurement types. Although psychologists will happily point out that many of these correlations are 'statistically significant', it is perhaps sobering to note that r^2 values of 'explained variance' span a range of .44 to as little as .03. Interestingly, these low observed correlations are by no means unique to this example, Beaudoin and Desrichard (2011) observed in their meta-analysis of 107 studies looking at the relationship between general memory self-report and actual memory performance, a correlation of .15, whilst Hendel and Brysbaert (2024) reported that the correlation between subjective and performance measures of emotion perception approached zero (i.e., −.02 − see also Murphy & Lillienfield, 2019).

All this serves to illustrate that any observed CFMT/PI20 correlation will have a considerable level of *uncertainty* around it, which naturally reflects varying degrees of *measurement error* and will impact any attempt to identify prospective DP research candidates (we will return to this topic in the General Discussion). The degree of this problem is illustrated by Fig. 1 below (inspired by Arizpe et al., 2019), where we plot a large general population distribution (N = 1344) on the two tests of note (this sample will be used in the current work and the correlation in this case is r = -.30, 95% CI [−.25, −.35]). In addition, two lines serve to highlight individuals with 'extreme' scores on each test: that is a PI20 score commensurate with very poor self-reported day-to-day experience, or a CFMT score indicating poor unfamiliar face memory performance. As a general convention, researchers select this boundary for 'extreme' performance based on observed scores that are at least two standard deviations from the population mean (likely because these observations lie outside a 95% confidence interval).

In this case (following Arizpe's example), we will use the PI20 as our 'predictor' such that we might assume that if the two measures are well aligned, all individuals who score themselves very highly on the PI20 (above the cut-off) will *also* perform very poorly (below cut-off) on the CFMT (i.e., fall into the box marked 'True positives': 5/1344), and the reverse should also be true (i.e., 'True negatives':1255/1344-people who say they are ok at face recognition and score similarly on the CFMT). Unfortunately, there are examples of 'misclassifications' of individuals, with some individuals possibly subjectively 'underestimating' their CFMT score (see top right-hand box-'False positives': 45/1344), or the *reverse* with individuals possibly subjectively 'overestimating' their CFMT score (i.e., fall into the box marked 'False negatives': 39/1344). This latter example may also be reflected by reports of some cases with prosopagnosia who, before accidental discovery, had *no prior* awareness that it was a face recognition problem that was causing their difficulties in social interaction (e.g., Susilo et al., 2010), or potentially that 'real-world' performance for such individual is generally good because of various forms of compensation, and so 'overestimation' on the CFMT follows. Consequently, it is apparent that the level of individual 'insight' into *any* given cognitive process, such as face recognition, likely lies on some continuum, and the Figure above provides another means of illustrating the low underlying correlation between the measures.

These examples thus highlight the unhappy state of current affairs for the field of researchers interested in studying individual differences in face processing ability, and in particular those interested in recruiting 'extreme' subpopulations. That is, the correlation between established, ostensible performance, and subjective measures is rather poor (in psychometric terms, there is poor *convergent validity*), and there is no clear reason for this. Notwithstanding the obvious likely issues of measurement error (which we would argue apply to *all* tests), there is the issue that a key problem for someone filling in a 'subjective' measure of their day-to-day face processing ability is that they must have a good meta-cognitive grasp of how they perform *relative* to the general population (see Kramer, McIntosh, & Nuhfer, 2024; Kramer & Tree, 2023). Put simply, if I am about average at face recognition, it's much harder to judge exactly how average, which implies that it might be fair to assume that it is perhaps easier to do so if I *know* I am an 'extreme' performer than otherwise. In other words, the individual's own day-to-day experience may help them *realise* their 'divergence' and thus this might lead to an elevated level of performance 'insight' (which is masked in the general population observations). However, singularly poor day-to-day experience may

not be sufficient to draw this inference. Given that individuals may be both poor performers *and* have generally low metacognitive insight, consistent with those we highlighted as 'False negatives' in Fig. 1 above (bottom left box). In such cases, individuals appear possibly 'overconfident' in their abilities. Though it remains unclear whether this 'overconfidence' stems from an overestimation of their *own* abilities or an underestimation of *others'* abilities since in many cases this is a *relative* judgement or reflects other issues such as the fact their 'day-to-day' ability may be boosted by various compensation strategies—an issue we will revisit in the General Discussion.

With these issues in mind, the work of Estudillo and Wong (2021) is relevant for examining the relationship between 'subjective' and 'objective' performance face-processing tasks, particularly when considering 'extreme' performance. In this case, the authors sought to investigate whether the level of insight varied across a population range of face recognition performance. To do so, they subdivided their sample into *performance quartiles* on the ability test (CFMT) and re-assessed the correlation with PI-20 scores per quartile (following in a similar tradition of 'insight' research on the 'overconfidence' effect, Dunning, 2011)). They found that a group–level correlation was statistically 'significant' for the lowest ($r = -.26$) and highest quartiles ($r = -.28$) *only*. They further replicated these findings in a secondary dataset reanalysed from the work of Gray et al. (2017, $N = 425$). They concluded that *only* people at the lowest or highest ends of actual ability have some (albeit clearly modest) metacognitive 'insight' into their level of performance. However, there has been considerable criticism of the approach of using quartiles in this manner in that it appears to create statistical artifacts that throw doubt on subsequent interpretations (see Kramer, McIntosh & Nuhfer, 2023). Namely, the problem of *regression to the mean*-i.e., dividing a group into 'quartiles' artificially exaggerates differences in self-assessment, creating the illusion of the Dunning-Kruger effect, even in *random data*; and *range restriction* which reflects the fact that analysing correlations within quartiles reduces statistical power, making it falsely appear that metacognitive insight exists only at the extreme. With that in mind, other researchers using *a trial-level insight* approach (i.e., responses to targets with associated 'confidence judgements'), have reported that individuals who performed objectively 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 and Dodson, 2024; Kramer, 2023; Kramer et al., 2022). Whereas poor performers' confidence ratings failed to differentiate between their correct and incorrect responses, suggesting poor insight (as measured by confidence) and performance are associated.

There thus remains an open question: might *some* individuals have elevated insight into how poor their face abilities might be (consistent with the 'True positives' in Fig. 1), and if so, how might one identify them? At this point, a salient example relates to how many research labs currently recruit DP research candidates (i.e., very poor face recognition ability). Individuals often *'self-refer'* to DP research labs, explaining that in their own experience, they have reason to believe they have very poor face recognition ability (they are 'self-identifying' as likely DP). Perhaps unsurprisingly, individuals who approach DP labs for research respond to subjective measures by clearly indicating the severity of their problems (i.e., they easily score at levels that meet PI20 'cut-offs'-see Fig. 1), and this led to

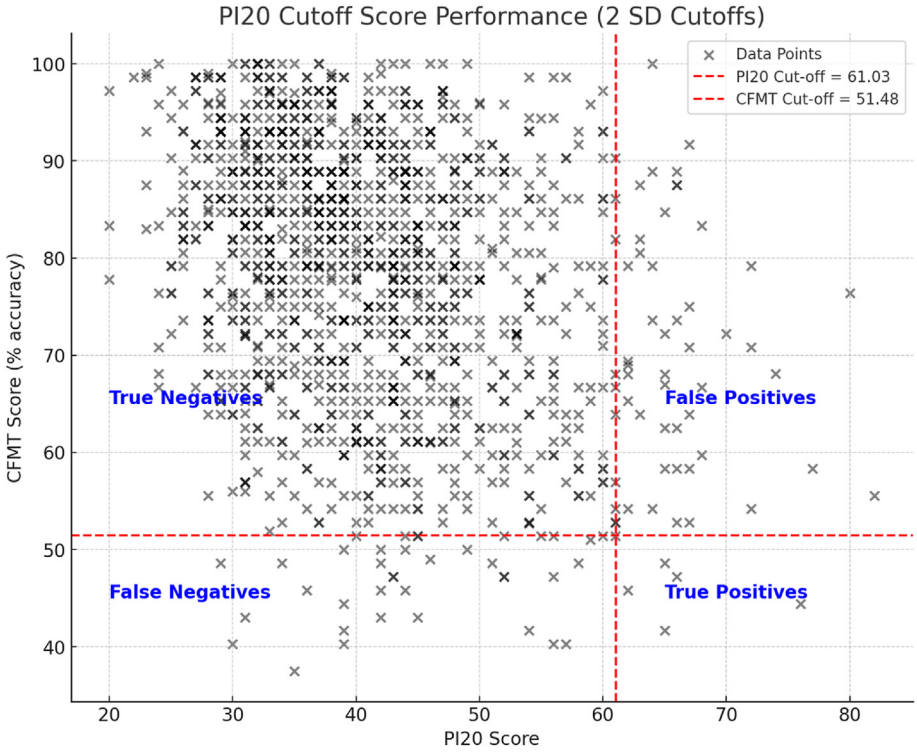

Fig. 1 – General population scores on CFMT/PI20 with 'extreme' (2SD) cut-offs.

them being sufficiently motivated to take part in research! In this case, there is a 'self-refer' subpopulation that comprises (by definition) individuals who are *above* the 'cut-off' line for the subjective face measure (PI20), *and their day-to-day experience is such that they 'self-identify' as DP* (which may not be true of all of those in the general population we might observe above the PI20 'cut-off' in Fig. 1 above). We then have a straightforward research question-might individuals who so 'self-refer' to DP research labs reflect a sub-population with *greater* insight than that of the general population? Put simply, might this group more obviously represent the individuals who fall into the 'True positives' sub-group presented in Fig. 1?

1.1. The current study-aims and objectives

The current work seeks to thus further explore the relationship between face processing ability and related 'insight' into that ability-if one assumes a degree of association between the objective and subjective measures of face recognition performance at a population level, one might ask, what about those who have 'self-referred' as having a problem as a subgroup? In other words, does this association between subjective/objective measures *interact* with group membership? As we have established, it is likely that across the general population not only does objective face performance vary, but so too does 'insight' into this performance (illustrated by Fig. 1 above). In the case of 'self-referrers'-who are individuals who so strongly *think* they have a problem that they are willing to spend time being tested-as a population do they disproportionately represent those on the high end of the 'insight' continuum? A second key aspect of the current work is to abandon the idea of 'cut-offs' (since we treat our 'self-referring' DPs as a potential sub-group), whilst also adopting a Bayesian statistical modelling framework. This approach has many benefits over the classical tests. By taking a model-based perspective, we can learn the associations between objective and subjective measures within control and 'self-referred' groups *simultaneously* and use the model to make counterfactual predictions about objective performance using any desired level of subjective ability and group assignment. The distinct advantage of adopting a Bayesian model is that these predictions are *distributions*, which allow for probabilistic statements about a range of hypotheses, as opposed to the probability of the data under the null hypothesis. We can also leverage credible intervals to simply state the probability estimates are within a certain bound, as opposed to confidence intervals (Kruschke, 2018). As such, we can make more flexible and informative contrasts between self-referred individuals and others.

In summary, the current study investigates a key question-do individuals who 'self-refer' to testing for developmental prosopagnosia (DP) show (as a subpopulation) elevated levels of 'insight' into their face recognition ability? As we have established, population-level correlation-based studies of objective/subjective performance will include individuals who exist across the continuum of insight ability and thus may not be particularly helpful in interpreting the key DP individuals for research. In our work, we side-step this issue entirely by treating the entire self-referral group as a distinct subpopulation to see if they do indeed have greater insight.

2. Methods

2.1. Participants

As we elaborated in the introduction, we are interested in the degree to which a specific population of individuals who 'self-refer' have a higher degree of insight into their face recognition abilities, as compared to the general population. To achieve this goal we obtained already published data sets focused on two types of key samples-on the one hand, a general population sample (i.e., individuals recruited via general UG populations, or online platforms such as Prolific) and on the other hand a sample of individuals who have approached a DP lab as volunteers saying they experience day-to-day challenges with face recognition ('self-referrers'). We discuss these and their specifics below.

2.2. Normative samples from the general population

With N = 1,344 in total, we gathered the following datasets from various existing published sources where participants were recruited across a number of sources. These include Burns et al. (2017) N = 48 recruited online (28 identified as males, 20 identified as females, mean age = 38.81, stdev = 11.33), Tsantani et al. (2021) N = 238 recruited online via Prolific (104 identified as males, 131 identified as females, mean age = 36.56, stdev = 11.72), Gray et al. (2017) N = 142 (sample a) undergraduates recruited at City University (56 identified as males, 86 identified as females, mean age = 29.22, SD = 11.91), Gray et al. (2017) (sample b) N = 283 undergraduates recruited at Reading University (101 identified as males, 180 identified as females, mean age = 26.64, SD = 13.16), Shah et al. (2015) N = 97 recruited via a local participant database (37 identified as males, 60 identified as females, mean age = 29.62, stdev = 13.11), Tagliente et al. (2023) N = 536 undergraduates recruited for testing online and in lab (200 identified as males, 334 identified as females, mean age = 27.42, stdev = 10.44).

2.3. Self-referring samples

We obtained key samples of 'self-referring' DPs all of whom were recruited online, with N = 279 in total. These include Burns et al. (2017) N = 61 (9 identified as males, 48 identified as females, mean age = 41.46, stdev = 14.02), Murray and Bate (2019) N = 47 (16 identified as males, 31 identified as females, mean age = 53.02, SD = 13.30), Tsantani et al. (2021) N = 146 (39 identified as males, 105 identified as females, mean age = 37.18, stdev = 10.72), Shah et al. (2015) N = 25 (16 identified as males, 9 identified as females, mean age = 45, SD = 17.70). All data is either publicly available or by request via the referenced papers discussed above, and it was immediately apparent that the normative sample is younger (Mean = 29.63, SD = 12.29 vs 41.49, SD = 13.86) and that the 'self-referral' group was disproportionately female represented (N = 80 identified as male, N = 193 identified as female), relative to the general population sample recruited (N = 526 identified as male, N = 811 identified as female). It would appear that in general, proportionately more female

Table 1 – Correlations between PI20 and CFMT scores across studies.

Study	Sample Description	CFMT Correlation (r)	R^2	N	Notes
Shah et al. (2015)	Local database	−.6683	.4466	87	Included DPs (without DPs, r = −.315)
Marscholek et al. (2019)	Social media (Polish)	−.4200	.1764	1270	
Ventura et al. (2018)	University students (Portuguese)	−.4300	.1849	123	
Gray et al. (2017) (sample a)	University students (London)	−.3940	.1552	142	
Gray et al. (2017) (sample b)	University students (Reading)	−.3900	.1521	283	
Estudillo and Wong (2021)	University students (Mandarin)	−.3500	.1225	255	
Tsantani et al. (2021)	Prolific	−.2670	.0713	238	
Stantic et al. (2021, Experiment 2)	University (online)	−.2610	.0681	97	
Stantic et al. (2021, Experiment 1)	University (lab)	−.2190	.0480	40	
Stantic et al. (2021, Experiment 3(T1))	University (online)	−.2190	.0480	68	
Stantic et al. (2021, Experiment 3(T2))	University (online)	−.1640	.0269	68	

volunteers tend to present at DP labs, in that more than *twice* as many women 'self-refer', though it remains unclear why that should be. Interestingly, recent work by DeGutis et al. (2023) suggests that 'insight' may well vary across gender (i.e., males tend to overestimate/females underestimate) and age (poorer 'insight' for much older participants)-and thus in our analyses we ensure that both gender and age are included. For our analyses, we obtained the raw scores of participants undertaking two key tests (described below) reported in many research studies and discussed earlier (i.e., CFMT and PI20).

2.3.1. *Test materials*

The samples selected were tested on two measures of face recognition-one 'objective' (the Cambridge Face Memory Test-CFMT) and one 'subjective' (the Twenty Item Prosopagnosia Index-PI-20). Both are routinely used in the screening of volunteer DP cases by a number of labs around the world and have typically high reported internal reliabilities based on Cronbach Alpha (Childs et al., 2021: CFMT a = .917, Shah et al., 2015: PI20 a = .84). Full details for the measures can be found in the key CFMT paper (Duchaine & Nakayama, 2006) and the key PI20 paper (Shah et al., 2015). In Tables 2 and 3 we outline the mean/standard deviation scores on both measures across the key groups we have selected. A somewhat obvious observation from these two Tables is that mean scores for the normative samples do vary somewhat, as do sample sizes-suffice it to say this will always have consequences on subsequent participant selection via things like z-score cut-offs (discussed in Fig. 1). In addition, as mentioned earlier, many studies typically report the correlation between the objective/subjective measures, and in the interests of similar transparency we do so here-for the normative sample the correlation between PI20 and CFMT was significant, $r(1344) = -.301$, $p = <.001$ (lower 95% CI = −.349, upper 95% CI = −.252), and this is consistent with the level of correlation reported in many other studies mentioned earlier (see Table 1). The correlation for the DP 'self-referral' group was also largely similar $r(279) = -.281$, $p = <.001$ (lower 95% CI = −.386, upper 95% CI = −.169). As a historical aside, in their original presentation of the PI20, Shah et al. (2015) criticised an example of the then-available subjective face measure (devised by Kennerknect et al., 2006), because it '*correlates poorly*' with objective face measures-they write *"Published correlations between scores on this scale and objective tests of face recognition ability range from*

Table 2 – CFMT accuracy Scores Across Different Population Samples.

Study Samples					
Normative Sample	N	Mean	StdDev	Min Score	Max Score
Burns (2024) norm	48	.829	.136	.403	1.0
Tsantani et al. (2021) norm	238	.74	.138	.403	.958
Gray et al. (2017) (sample a) norm	142	.807	.128	.458	1.0
Gray et al. (2017) (sample b) norm	283	.768	.129	.472	1.0
Shah et al. (2015) norm	97	.807	.129	.49	1.0
Tagliente et al. (2023) norm	536	.799	.134	.375	1.0
DP Sample					
Burns (2024) DP	61	.608	.136	.292	.931
Tsantani et al. (2021) DP	146	.602	.122	.319	.931
Murphy & Lilienfeld (2019) DP	47	.504	.062	.361	.667
Shah et al. (2015) DP	25	.566	.097	.28	.72

Table 3 – PI20 scores across different population samples.

Study Participants					
Normative sample	N	Mean	StdDev	Min Score	Max Score
Burns (2024) norm	48	39.08	8.97	20	62
Tsantani et al. (2021) norm	238	44.85	10.7	23	80
Gray et al. (2017) (sample a) norm	142	40.11	9.58	23	68
Gray et al. (2017) (sample b) norm	283	41.69	10.07	20	74
Shah et al. (2015) norm	97	38.52	9.23	23	65
Tagliente et al. (2023) norm	536	40.59	9.21	22	82
DP sample					
Burns (2024) DP	61	81.87	7.64	62	95
Tsantani et al. (2021) DP	146	77.72	7.30	65	96
Murphy & Lilienfeld (2019) DP	47	81.51	6.87	67	99
Shah et al. (2015) DP	25	79.72	9.9	62	97

r = .20 to .55", and yet sadly with the benefit of time and hindsight it appears this "*low association*" mirrors what we (and others) observe for the PI20. As a consequence, it is not apparent that a different set of subjective questions necessarily improves subjective/objective correlations in this case unfortunately (see Tree, 2011 for an early discussion).

2.3.2. *Analytic strategy*

To more fully interpret this data, we used model-based Bayesian inference, specifically fitting a hierarchical linear regression to the full dataset. Z-scored CFMT scores were predicted from Z-scored PI20 scores (continuous measures), a dummy-coded variable indicating group membership (the normative samples coded as zero, and the self-refer samples coded as one), and the interaction between these two variables. Additionally, participant sex was entered as a covariate (dummy coded with female as the reference category, and categorical predictors for male, nonbinary, trans male, other, and non-disclosed, as reported by participants) alongside Z-scored participant age. Five observations were dropped due to missing data for these covariates, leaving a final sample of 1,618. The random effects included an intercept for each dataset and a dataset-specific slope for the PI20 effect. This parameterisation allows the model to account for variation in the association between the CFMT and PI20 in each dataset, preventing, for example, any one dataset with a particularly low or high association from driving the results, or for the possibility of range restrictions in some datasets biasing associations in a specific direction. This specification induces partial pooling, which gives a more robust estimate of the association between the CFMT and PI20 across all datasets (Gelman & Pardoe, 2004). As an alternative, we could ignore the variation across datasets (i.e., complete pooling), which might open the analysis to bias-larger datasets could drive the overall pattern and remove important variation present in some datasets but not others. Conversely, we could fit many models-one for each dataset-which would return many estimates, but with an unclear overall association, and with the individual estimates having poorer precision given that each dataset is necessarily smaller than all of them combined.

As our model is Bayesian, we must specify prior distributions on the parameters. We set weakly informative priors on parameters (Gelman et al., 2020), specifically so they had little influence on the data. We used a Gaussian likelihood (reflecting an assumption that CFMT scores are normally distributed). For the intercept and coefficients of age, sex, PI20, group, and their interaction, a Gaussian distribution with a mean of zero and a standard deviation of ten was used, which entertains very large effects in either direction. A half-Gaussian distribution with a standard deviation of four was used for the error variance of the likelihood. For both of the random effects of the dataset (intercept and PI20 slope), a Gaussian distribution with a sum-to-zero constraint was used. The standard deviation of these distributions had a hyperprior of an Inverse-Gamma distribution with a mean and standard deviation of one. This hierarchical approach allows us to estimate the variability in the dataset's random effects. A formal statement of our model is as follows:

$$\begin{aligned}\text{Z-CFMT}_{ij} = {} & \beta 0_j + \text{Z-PI20}_j * \beta_{1j} + \text{Group} * \beta_2 + \text{Z-PI20} : \text{Group} * \beta_3 \\ & + \text{Z-Age} * \beta_4 + \text{Sex} * \beta_5\end{aligned}$$

Where i indicates an observation, and j indicates a dataset.

Models were estimated with the PyMC package (Salvatier et al., 2016) in the Python programming language. Four Markov Monte Carlo chains were run, with each having 2,000 tuning steps and 5,000 samples drawn from the posterior. The model converged, and all parameters had an $R = 1$.

2.3.3. *Model interpretation*

The interaction between PI20 and group position (normative of self-referring) directly tests whether the association between PI20 and CFMT scores is different within each group. We recovered the slopes for each group by adding the interaction coefficient to the PI20 slope and used the interaction coefficient itself as the difference between the groups. We examined the posterior distribution of these effects to draw inferences about the hypothesis that the groups differ in their insight into objective performance, calculating the mean and 94% highest-density intervals (HDI), to show the expected and credible range of effects, respectively. We also calculated the posterior probability of direction (Makowski et al., 2019), similar to a frequentist *P*-value, to discern the direction of the effect that was strongest. This is easily calculated by examining the proportion of the posterior distribution above or below zero, given the observed data. Note that this value indicates the probability of the effect is greater than zero, given the data, and not the converse, as in frequentist approaches (Welsch et al., 2023).

3. Results

After estimating the model, we predicted the expected CFMT Z-score across the range of Z-scored PI20 values, for both the normative and self-referral groups, holding constant age, sex, and the random effects. These predictions are illustrated in Fig. 2, showing the pattern of responses across the data-the estimated associations show clear and substantial overlap between groups. Overall, the model explained around 34.5 % of the variance in CFMT scores (94 % credible interval [31.5%, 37.3%]. The estimates of the slopes for each group are also shown in Fig. 2. For the normative sample, the slope was negative, $b = -.47$, [−.64, −.30], $p(\theta < 0) = 100$ %, and for the self-refer group, it was similar in magnitude, $b = -.48$, [−.79, −.18], $p(\theta < 0) = 99.7\%$. Importantly however, the interaction term-that is the *difference* between these slopes-centred almost on zero, $b = .011$ [−.44, .4], $p(\theta > 0) = 47.7$ %. That is, the model suggests a probability of just under 50 % that self-referring participants have a higher association between their PI20 and CFMT scores than the normative sample, the slope being on average just .01 units higher.

As our model is Bayesian, we explored the full range of credible differences between the groups that are consistent with the data. The bottom panel of Fig. 2 illustrates this. The model predicted the expected CFMT Z-score for a wide range of PI20 Z-scores for both a normative and self-referring group, holding the covariates constant. At each PI20 Z-score we

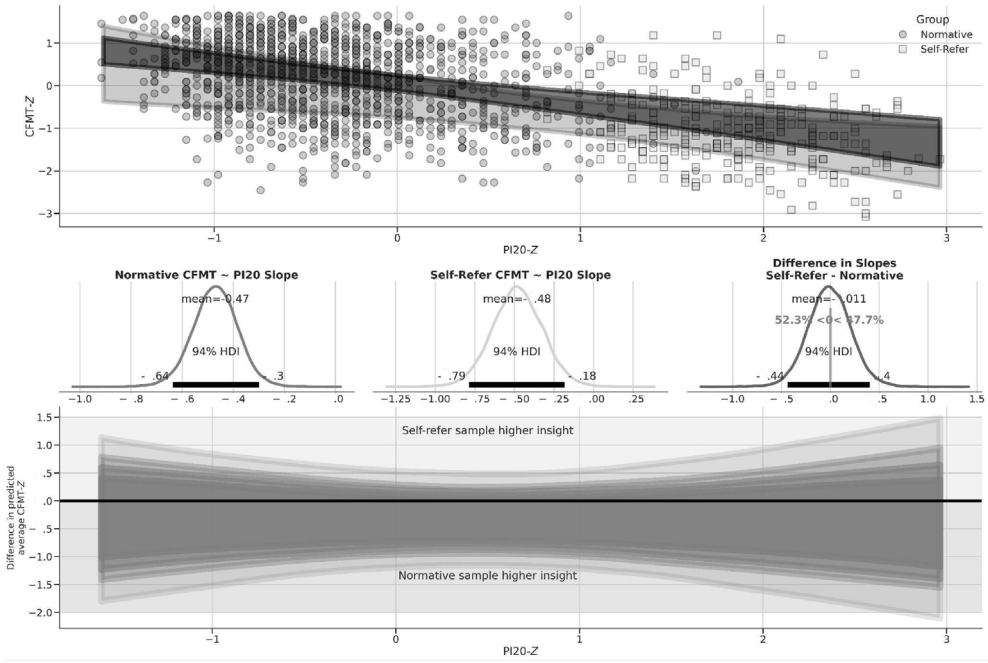

Fig. 2 – Top row-the predictions of the model over the raw data. Shaded bands indicate the 94% credible intervals. Middle row-the posterior distributions of the PI20 slope for each group, and the difference in those slopes. Bottom row-the posterior distribution of the differences between both groups evaluated across a wide range of PI20 scores, highlighting that a small majority of this difference is in the direction of the normative sample.

computed the difference between these estimates, subtracting the normative group from the self-referral group. These estimates are distributions, and we computed varying credible intervals from them (80, 90, 95, 99, increasing in lightness in Fig. 2). Plotting this shows the interaction effect from a different angle-across the range of PI20 scores, there is no real difference between the groups in terms of their CFMT score-if anything, most of the posterior mass of these differences are in the direction of the normative sample having higher CFMT scores. For example, consider a pair of individuals, one from the normative and self-referring group, scoring a +2.96 Z-score on the PI20 (the highest observed PI20 score in the dataset). The probability the self-referring individual has a lower CFMT score than the normative individual, reflecting their concerns, is 27%. Conversely, an individual with a PI20 Z-score of .81, the lowest observed in the self-refer group-has a probability of 13% of having a lower CFMT score. Note that these probabilities are veridical; the complement is the probability the self-referring individual has a higher CFMT score.

4. General Discussion

The current study sought to further explore the relationship between participant's ratings of their own day-to-day face processing ability (PI20) as a predictor of performance on a key 'objective' unfamiliar face memory task (CFMT)-namely the degree of 'insight' individuals might have in their face processing ability. Previous correlational-based studies suggested only a modest association, but these were often based on general population samples. Here, we tested whether individuals who come forward with concerns about their facial recognition abilities, willingly volunteering their time for research, do indeed have greater 'insight' as compared to the wider population-consistent with this observation, Tsantani and colleagues point out such individuals *"frequently travel long distances with little or no compensation and are typically conscientious, committed participants."* (Tsantani et al., 2021). It thus seems plausible to assume that the association between predicted face performance (PI20) and actual face performance (CFMT) should be better for those who 'self-identify' as DP, for two reasons: (a) those with very poor performance might expect better *relative* insight-stated another way, if an average performer is asked whether they are above or below average, this should be a harder relative judgment than if their performance where at the extremes of the distribution (discussed more below), and (b) people willing to volunteer their time, clearly perceive they have a problem (and can show significant distress about it) and these concerns should count for something (see Burns, Gaunt, Kidane, Hunter, & Pulford, 2023).

However, the available data suggests there is little evidence that 'insight' is indeed higher in this 'self-referred' DP population. The Bayesian model used here, testing the interaction between group status and PI20 score, showed no credible evidence of a difference between the groups. In effect the same modest levels of predictive performance are seen in this group as that seen across the wider population, consistent with the many correlational study findings discussed earlier. These low correlations could reflect: (a) the fact that face processing 'insight' is indeed generally poor or (b) that the low correlation has been '*attenuated*' by high levels of measurement noise

(Cooper, 2023; Spearman, 1904). This latter problem is not unique to the 'subjective/objective' face processing comparison, since similarly low correlations across *different objective* face processing tasks have also been reported (e.g., Bobak et al., 2023; Fysh & Ramon, 2022). In the next few sections, we discuss some key issues facing the field at present (relating to (a) and (b)), with key observations and suggestions for future approaches to make forward progress.

4.1. *The dangers of noisy observations in experimental design and 'objective' measures*

The observation of these poor cross-task correlations—and their implications of poor test *reliability* (discussed later)—has consequences for other studies often undertaken with DP populations, using DP/Control as an "independent variable" in an ANOVA (or t-test) design. In this case, "group" is not a true experimental independent variable because it cannot be freely manipulated (and thus randomized); rather, it reflects an inherent property of the *participants* (i.e., high/low face performers). Brysbaert (2024) makes this point using the example of Woodhead and Baddeley (1981), who compared a group of people who were good at memorizing faces with a group who were poor at it. They found that the good group was also better at remembering paintings but not at remembering words—suggesting a dissociation between visuospatial and verbal material. Brysbaert points out that a critical issue was that:

> *"Woodhead and Baddeley (1981) were not able to randomly place people in the condition of good and bad face recognition. All they could do was select people based on an existing difference, making their design a correlational design even though the data were analyzed with analysis of variance. The findings of the study are best summarized by saying that there was a correlation between memory for faces and paintings, but not between memory for faces and words."*

Critically, if the key correlation is already low, then small-sample comparisons using this approach will be vulnerable to *uncertainty* because of measurement noise. This vulnerability is nicely reflected in simulations provided by Brysbaert (2024; see Fig. 2), where a small correlation of .20 can be misjudged as much higher or much lower (e.g., −.6 to +.7) if sample sizes are too small. By contrast, the effective confidence interval shrinks (e.g., +.1 to +.3) with much larger samples (the author suggests N > 400). Moreover, this also illustrates that "power" simply improves one's confidence in increasingly *marginal effects not being zero* (as mentioned earlier, if r = .2, then r2 = .04). Work in classical cognitive neuropsychology has often recognized this (see Nickels et al., 2011) and is thus reluctant to treat "groups" as representative patient samples; instead, it typically uses *case-series* approaches (see Wingrove & Tree, 2024 for a recent discussion). With respect to the current work, by employing a hierarchical model we are able to reduce the influence of noisy measurements in each of the individual datasets; through partial pooling, the model we used estimated an overall association between the PI20 and the CFMT as well as the individual associations in each dataset, but with estimates of smaller datasets being pulled toward the overarching effect, leading to a more robust model. However, this does not affect the magnitude of the association, only helping with the stability of the estimate.

Unfortunately, some researchers have assumed they might escape the issues posed by low observed correlations between the PI20 and CFMT at the population level by adopting exactly the kind of group approach described above by Brysbaert (2024). An anonymous reviewer brought this to our attention with the example of Tsantani et al. (2021). These authors suggested that the low observed correlation in the general population (as shown in our Table 1) *"does not mean that the PI20 is ineffective at distinguishing likely DPs from likely non-DPs (i.e., its intended purpose)."* We appreciate the reviewer raising this point because it directly relates to the question we investigated in the current study—namely, whether insight is indeed higher for self-referred DPs-and also relates to our earlier point about using a categorical-group treatment, since this is precisely the approach Tsantani and colleagues adopted to investigate that same question. As they put it: *"In keeping with its [PI20] intended use, we adopt a group design (not a correlational approach). We use the PI20 to identify two groups of participants: suspected DPs (high scorers) and suspected non-DPs (low scorers). Having defined groups of participants based solely on the individuals' PI20 scores, we examine how these groups differ in their performance on objective measures of face recognition ability (two variants of the CFMT)."*

For context, Fig. 1 (above) illustrates Tsantani and colleagues' method: they separated the population into two groups based on a PI20 "cutoff" (equivalent to the left and right of the vertical line in Fig. 1) and following ANOVA, they performed a t-test on their average CFMT scores. Indeed, when Tsantani et al. (2021) used a cutoff of 65 on the PI20 (slightly higher than shown in Fig. 1), the t-test yielded a statistically significant group difference for CFMT performance across two versions of the test (effect sizes of 1.085 and .88). They interpreted this finding, in contrast to the often-low observed population–level correlations (see Table 1), as follows: *"Critically, the PI20 is a measure of prosopagnosic traits, not of face recognition per se. The scale has little ability to distinguish people who are slightly below average, from those slightly above average, from super-recognizers. As such, the modest correlations described above are uninformative about the validity of the scale. Insofar as correlational approaches assume a* ***linear relationship*** *between PI20 score and CFMT performance across the entire range of abilities, they are ill-suited to the validation of the PI20. The group design used here—in particular,* ***the categorical treatment of anyone who scores below cut-off as unimpaired****—provides a fairer test of the validity of this instrument."* (bold for emphasis).

However, as we have explained, this group-based approach is not truly an *alternative* to a correlational approach; it is essentially a recasting of it—and the observed t-test results are *entirely consistent* with such a linear relationship between the PI20 and CFMT. Indeed, the significant group difference merely reflects the same underlying correlation. Put simply, finding a significant group difference does not imply that the PI20 is more "meaningful" when used categorically. It is easy to see why confusion can arise when a continuous measure (like the PI20) is artificially split into "high" vs. "low" groups, but doing so still represents a correlational design and, as Brysbaert (2024) explains in detail, can even reduce statistical

power. We can illustrate this point by using a linear regression and showing that predicted differences across "groups" are consistently observed at *any* "cutoff" on the PI20 (this is akin to sliding the vertical line along the x-axis in Fig. 1).

We can consider a simplified example with these data in the current study, ignoring its hierarchical structure and relying on simpler frequentist point estimates. We split the dataset into two groups, as suggested by Tsantani et al. (2021)-anyone with a PI20 score of 65 or above is a 'high' scorer, and 'low' otherwise. A *t*-test comparing CFMT scores for those groups (high-$M = 59.17$, low-$M = 78.74$) suggests a clear and significant mean difference, $M = 19.57$, $p < .001$, which could be taken as evidence that the PI20 accurately classifies poor objective performance when used as a classifier. However, consider a simple linear model of these two continuous, untransformed measures-a simple linear regression predicting the CFMT from the PI20 suggests an intercept (mean CFMT score when the PI20 is zero) of 98.58, $P < .001$, and a slope (change in CFMT score when the PI20 increases by one point) of -.49, $P < .001$. By taking the predictions of this model-the expected values of the CFMT given a PI20 score-and averaging them above and below the cut-point, we find an average high-group prediction of 59.96 and a low-group of 78.55, with a clear and significant mean difference in predictions of 18.59 units-practically identical to the observed mean difference induced by cutting the PI20. A surprising consequence of this linear model is that it does not matter where we place the cutpoint, the mean difference will be relatively consistent. Consider a cut point of 50 that separates the groups-now the 'low' group has a mean CFMT score of 79.94 and the 'high' group a mean score of 64.67, a significant difference of $M = 15.27$ units. Equivalently, the model's predictions suggest a mean predicted difference of $M = 15.43$ units. A further consequence of this model is that simply by standardising both variables we obtain the correlation between them, $b = -.56$, similar in magnitude to that observed by Tsantani et al. (2021).

In summary, it is clear that underlying cross–task correlations are *always* key to the study of individual differences research and cannot be simply escaped (at least in this specific case) by chopping up observed population distributions into 'groups' via 'cutoffs' as illustrated in the example above. To reiterate, our own work sought to side step these issues by classifying 'groups' in a more independent fashion (namely 'self-referrers' vs. the general population) and adopting a hierarchical model rather than a straightforward linear regression approach-nonetheless our findings mirror those of Tsantani et al. (2021) in that we find no evidence of a higher 'insight' (or interaction with observed CFMT score) for the DPs.

Importantly, the implications of observed poor cross task correlations extend beyond group differences studies to observations of *individuals* themselves. In this case, the field of psychometrics has been aware of the consequences of poor correlations on the *reliability* of measures (see Cooper, 2023), in particular on the pernicious issue of *regression to the mean*. In the convention of classical test theory, any individual observed score comprises the *true score* plus some level of measurement noise, which can be captured in the *reliability* of the measure used. That is, one would want to assume that the measurement of the individual is *stable* in some sense and thus reflects a 'trait' (or 'latent' ability) of that individual rather than some transient 'state' they might have been in at the time of measurement (like mood). Many cognitive psychologists often implicitly make this former assumption in the paradigms they use without any actual evidence of its ground truth. Currently, there is a great deal of lively debate about these issues for studying individual differences in cognitive research (see Hedge et al., 2018; Rouder & Haaf, 2019; Satchell et al., 2023 as excellent examples), which is often overlooked in this case.

But let us provide a short illustrative example of the implications of what is at stake. A typical way of determining the 'stability' of an observed individual score is to either obtain another observed score on the same test again (Time 1 vs Time 2-T1/T2) or on another very similar (and highly correlated) task-that is, obtain data on *test-retest reliability* (Brysbart, 2024; Cronbach, 1947). In this case, if one observed an individual's extremely poor score on the CFMT, and we want to interpret such a score as reflecting the 'stable' fact that this individual is indeed 'extremely bad' (i.e., DP), it naturally follows that they should be expected to do equivalently poorly *if we tested them again*. For any 'diagnostic' approach this key assumption is apparent; for example, Degutis et al. (2023a) (see also Burns et al., 2023, 2024) have suggested best practice might be to "*adopt standardized neurocognitive disorder cutoffs from DSM-5 to identify major (self-report + at least 2 validated face recognition tests z-score < -2) and mild (self-report + at least 2 validated face recognition tests z-score < -1) forms of prosopagnosia until more mechanistically grounded cutoffs can be identified.*".[1] Although this initially seems sensible, it is *entirely* dependent on the assumption that our observations of individuals are indeed 'stable' and reflect the *true score* (latent ability) of that individual. Unfortunately, it should be apparent that the cross task and test-retest reliability correlations key to this assumption suggest a less than satisfactory conclusion (see also Fysh & Ramon, 2022), given even well validated measures such as the CFMT and PI20 have less than perfect test-retest correlations being reported (CFMT-.68 (Murray & Bate, 2020) PI-20-.89 (Stantic et al., 2021). Critically, poor test-retest (or cross-task) correlations and the issues of regression to the mean work hand in hand to the extent that they will naturally *attenuate* any 'extreme' observed score to be most likely 'normal' (i.e., near the mean) on any subsequent observation (see Campbell & Kenny, 2002 for an excellent primer on this issue), which can easily account for reports illustrating individuals with 'extreme' scores on one measurement observation are most often no longer so on some other measurement observation (e.g., Bobak et al., 2023; Fysh & Ramon, 2022). Our own recent work on classifying candidates for 'other ethnicity blindness' discusses these dangers and suggests future directions for analysis of observed cases of 'extreme' performance (Tree & Jones, 2025).

[1] An anonymous reviewer suggested we report the number of our sample of 'self-referred' DP cases (279) who scored above a 'cutoff' of the PI20 of 65 (277/279) and of these how many scored below -2SDs on the CFMT (81/277) or -1SDs on the CFMT (204/277) so we provide this information here. Of course, this is entirely consistent with the low correlation and reliability issues which we subsequently discuss and thus Degutis and colleagues have suggested *multiple test* observations for objective performance.

However, we want to make clear that the interpretation of these cross–task correlations is not strictly that they mean individual differences research cannot be undertaken with them. Poor correlations can mask a variety of underlying issues, most salient of which may be that in all these cases we are correlating the *sum of scores* on any given measure, which is very likely to sharpen the impact of *attenuation* on those observed correlations (see Rouder & Haaf, 2019; McNeish & Wolf, 2020). Psychometricians have long been aware of the issues in this approach, with Spearman (1904) providing useful advice for attenuation adjustment using a well-established conventional formula. In addition, modern psychometrics has developed the item response theory approach to enable the calculation of *scaled scores* of test performance that more realistically reflect 'latent ability'-which we will be aiming to explore in future, particularly given in the case of CFMT, these parameters are already published (Cho et al., 2016). But our simple advice at this juncture is to point to the mature field of psychometrics which can provide useful pointers for studying individual differences in cognitive psychology (Brysbaert, 2024). We would also make two further additional pleas to the field in the future: (a) instead of providing summed scores on the tests we use, also provide scores by *individual items* and (b) if group differences remain the experimental design you wish to follow, start using a linear (mixed-effects) analysis approach so that participant and item level random effects are recognised. A straightforward (and hard-won) lesson learnt by both these authors is that 'collapsing' observed scores, either by averaging or summing, clearly removes critical variance relevant to the attenuation problem, and the approach should be avoided (see also Rouder et al., 2023).

4.2. *Subjective measures of day-to-day face processing ability-a heterogenous measure?*

In this study, we have used a well-established subjective measure of cognitive performance, the PI20, developed by Shah et al. (2015). It comprises 20 questions that tap a variety of different aspects of face processing experience. Inspection of these questions reveals such differences-e.g., *"My face processing is worse than most people", "I find it easy to picture individuals faces in my mind", "Anxiety about face recognition has led me to avoid certain social or professional situations", "I feel like I frequently offend people by not recognising who they are", "It is hard to recognise familiar people when I meet them out of context"*. As a consequence, perhaps part of the problem of correlating such a questionnaire with any particular experimentally based 'objective' measure is that some questions may be more or less relevant to performance on such a task, and thus likely add to the predictive 'noise' of the situation (see Kramer & Tree, 2024). In the field of psychometrics, it is common, when developing such subjective report measures, to undertake extensive factor analyses of the questions used to determine likely different underlying 'constructs'. In this case, Shah et al. (2015) did undertake an exploratory factor analysis (with varimax rotation), reporting a single factor that accounted for 61% of the variance. However, the sample used in this factor analysis comprised both controls and a large number of DPs (extreme performers) described above (see Tables 2 and 3). Unfortunately, since oversampling of extreme performers can influence factor structures, it is unclear whether the reported single-factor solution accurately reflects the measure's structure in the more typical general population. In any case, no subsequent *confirmatory factor analysis* has since been undertaken for the PI20 (however see Nørkær et al., 2023 who reported a CFA that was *not* consistent with a single-factor solution for a Danish PI20). As a consequence, we would suggest more work that follows the typical psychometric approach is undertaken with key subjective measures of face processing (such as the PI20) to understand the degree to which particular *questions* in such questionnaires may be better suited to 'predict' objective performance on different face processing tasks, such as unfamiliar face recognition, unfamiliar face matching, famous face recognition, etc., all of which comprise different kinds of tests used in DP labs. A good approach in this case, after confirmatory factor analysis, would be using item response theory analysis to identify key levels of question 'discrimination' for particular task performance for each case. This is very much in line with our recommendations to focus on disaggregated, item-level data analysis.

4.3. *Making subjective judgements of your objective ability is hard*

As mentioned earlier, regardless of the potential impact of different types of 'noise' on our observed scores—whether on subjective and/or objective measures—that may cause our data to deviate from the 'true' latent ability, it's important to acknowledge that assessing one's *own* performance on a cognitive task is inherently challenging. This is because it requires one to carefully calibrate one's ability compared to *others* (as mentioned earlier it can often comprise a *relative judgement*), when one is often left to guess what the normal distribution of task ability might *actually be* for the rest of the human race! Day-to-day face processing ability is an interesting example since it is fair to say the most likely 'naive' general view is that human face processing is remarkably good (it is a 'class apart'), such that the experience of a 'failure' must be rare. However, early work using diary measures suggests these mistakes may be more common in the general population than belies this 'naive' general assumption. Young et al. (1985) conducted a study where 22 participants recorded errors in face recognition over seven weeks, excluding the first week to allow for familiarisation with the recording process. A total of 922 incidents were documented, with no single participant contributing more than 18% of the errors; four major types of errors were reported with familiar faces: (a) person unrecognised, (b) person misidentified, (c) person 'seemed familiar' and (d) difficulty in retrieving full details of the person (older readers will certainly be familiar with *all* of these). As a consequence, it is apparent that despite the likely 'naive bias' of assuming the rarity of such episodes, real-world 'failures' are likely quite a common experience. In which case, individuals experiencing such episodes (which it seems is true of most of us), and then subsequently filling in a PI20 questionnaire, may also differ in the degree to which they might *remember* them, and if they do remember them how they *interpret them*. Interestingly, although this may apply to the consequences of 'day-to-day' experience on self-judgement, there is little

evidence that this necessarily plays a part for specific task exposure. For example, Tagliente et al. (2023) tested Italian participants where they specifically manipulated the order of administration of PI20/CFMT and reported that correlations were near identical (see also Gray et al., 2017 for a similar finding). However, other fields interested in subjective judgement and objective performance such as voice recognition ability have noted that correlations are often near zero but *can* improve with some task experience (Skuk & Schweinberger, 2013). In any case, we would argue some effort must be made to consider 'real-world' experiences and potential consequences on individual subjective reports of day-to-day face processing ability (see also Young & Burton, 2018 for a similar and more recent commentary).

Put simply, different individuals likely vary in their 'sensitivity' to poor performance, that is, perhaps like the subjective experience of other 'objective' states such as pain (Coghill et al., 2003), individuals' 'sensitivity' to this experience may vary considerably. In any case, it is likely individual responses to subjective measures such as the PI20 may well be impacted by such variables as social anxiety, low self-esteem, or personality traits such as neuroticism (see Megreya & Bindemann, 2013; Turano & Viggiano, 2017), *in addition to* variability in 'meta-cognitive insight' (i.e., knowing when one has made good or bad responses). The interactions between these variables remain an open empirical question. However, a recent excellent study by Grabman and Dodson (2024) is relevant to this discussion. These authors examined individual differences in face processing ability that used various measures of how people subjectively judge their own performance at the *trial level* (confidence ratings) and at the more general day-to-day face ability level (akin to the PI20). Critically, they found that this latter measure was linked to generally higher/lower levels of confidence in responses ('meta-cognitive bias') *regardless of trial level accuracy*. In other words, participants who subjectively rated their general face processing ability very favourably or otherwise, *also* had a trial level 'bias' to consistently respond at the upper or lower end of the confidence rating scale. This even affected the type of language participants used (i.e., the tendency to write phrases like '*extremely confident*'). Interestingly, this may well indicate a '*confidence trait*' across individuals (Kleitman & Stankov, 2007; Pallier et al., 2002, Stankiv et al., 2012), for which it would be useful to have some understanding of when interpreting responses to measures like the PI20 (and can explain our earlier observation of both 'overconfident' and 'underconfident' individuals in Fig. 1). Moreover, it would be interesting to determine the degree to which observed PI20 scores might be 'corrected' for this 'confidence trait' dimension (at either end of its continuum), and whether it might be a domain-specific (i.e., just faces) or domain-general (i.e., judgements of cognitive ability of any kind) phenomena-it may well also be linked to the observed gender differences in those who may 'self-refer' as potential DP candidates, a finding evident in the collated sample of data used here. These issues would also be relevant to other contexts of face processing research such as eyewitness memory where similarly researchers are interested in understanding the relationship between accuracy and confidence (e.g., Sporer et al., 1995; Wixted & Wells, 2017).

4.4. *Future directions for individual differences in face processing research*

The above discussion has highlighted several key challenges that face work that seeks to explore individual differences in face processing ability, whether measured by 'objective' or 'subjective' means. Our recent examples indicate that even individuals who 'self-identify' as DP appear as a group to have little improved 'insight' into their subsequent performance on a standardised 'objective' measure of unfamiliar face201 recognition (see also Burns et al., 2014), that is there is no credible evidence of higher levels of correlation in this specific example. We have discussed the difficulties with interpreting such observations of poor cross task correlations, which are entirely consistent with the converging evidence of several recent studies in this particular field (e.g., Bobak et al., 2023; Fysh & Ramon, 2022). However, all is not lost - we have made several suggestions for potential future work and urge the much wider utilisation of approaches that are very much standard practice in the field of psychometrics. Such approaches can both improve our understanding and interpretation of observed scores on many of our critical tests and provide a route forward to improving their sensitivity and utility. To a great extent our observations echo those of others in different fields of cognitive psychology (e.g., Brysbart, 2024; Hedge et al., 2018; Rouder & Haaf, 2019), and we hope may usher in a new era of exploring the interesting nature of individual variability in measures of cognitive performance.

Scientific transparency statement

DATA: All raw and processed data supporting this research are publicly available: https://osf.io/e8k6z/
CODE: All analysis code supporting this research is publicly available: https://osf.io/e8k6z/
MATERIALS: This research did not make use of any materials to generate or acquire data.
DESIGN: This article reports, for all studies, how the author(s) determined all sample sizes, all data exclusions, all data inclusion and exclusion criteria, and whether inclusion and exclusion criteria were established prior to data analysis.
PRE-REGISTRATION: No part of the study procedures was pre-registered in a time-stamped, institutional registry prior to the research being conducted. No part of the analysis plans was pre-registered in a time-stamped, institutional registry prior to the research being conducted.
For full details, see the *Scientific Transparency Report* in the supplementary data to the online version of this article.

CRediT authorship contribution statement

Jeremy J. Tree: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Conceptualization. **Alex L. Jones:** Writing – review & editing, Project administration, Investigation, Formal analysis, Data curation.

Code and data availability

osf.io/e8k6z.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

None.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cortex.2025.03.009.

REFERENCES

Arizpe, J. M., Saad, E., Douglas, A. O., Germine, L., Wilmer, J. B., & DeGutis, J. M. (2019). Self-reported face recognition is highly valid, but alone is not highly discriminative of prosopagnosia-level performance on objective assessments. *Behavior Research Methods, 51*, 1102–1116.

Bate, S., Bennetts, R. J., Gregory, N., Tree, J. J., Murray, E., Adams, A., Bobak, A. K., Penton, T., Yang, T., & Banissy, M. J. (2019b). Objective patterns of face recognition deficits in 165 adults with self-reported developmental prosopagnosia. *Brain Sciences, 9*, 133.

Bate, S., Bennetts, R. J., Tree, J. J., Adams, A., & Murray, E. (2019a). The domain-specificity of face matching impairments in 40 cases of developmental prosopagnosia. *Cognition, 192*, Article 104031.

Bate, S., & Tree, J. J. (2017). The definition and diagnosis of developmental prosopagnosia. *The Quarterly Journal of Experimental Psychology: QJEP, 70*(2), 193–200.

Beaudoin, M., & Desrichard, O. (2011). Are memory self-efficacy and memory performance related? A meta-analysis. *Psychological bulletin, 137*(2), 211.

Bennetts, R. J., Gregory, N., Tree, J. J., Banissy, M., Murray, E., Adams, A., Penton, T., & Bates, S. (2022). Featural and holistic processing can be separably impaired in disorders of face recognition evidence for two subtypes of developmental prosopagnosia. *Neuropsychologia, 174*, Article 108332.

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

Bobak, A. K., Jones, A. L., Hilker, Z., Mestry, N., Bate, S., & Hancock, P. J. (2023). Data-driven studies in face identity processing rely on the quality of the tests and data sets. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 166*, 348–364.

Brysbaert, M. (2024). Designing and evaluating tasks to measure individual differences in experimental psychology: A tutorial. *Cognitive Research: Principles and Implications, 9*(1), 11.

Burns, E. J. (2024). Improving the DSM-5 approach to cognitive impairment: Developmental prosopagnosia reveals the need for tailored diagnoses. *Behavior Research Methods*, 1–20.

Burns, E. J., Bennetts, R. J., Bate, S., Wright, V. C., Weidemann, C. T., & Tree, J. J. (2017). Intact word processing in developmental prosopagnosia. *Scientific Reports, 7*(1), 1683.

Burns, E. J., Gaunt, E., Kidane, B., Hunter, L., & Pulford, J. (2023). A new approach to diagnosing and researching developmental prosopagnosia: Excluded cases are impaired too. *Behavior Research Methods, 55*(8), 4291–4314.

Burns, E. J., Tree, J. J., & Weidemann, C. T. (2014). Recognition memory in developmental prosopagnosia: Electrophysiological evidence for abnormal routes to face recognition. *Frontiers in human neuroscience, 8*, 622.

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

Childs, J., Jones, A., Thwaites, P., Zdravkovic, S., Thorley, C., Suzuki, A., Shen, R., Ding, Q., Burns, E., Xu, H., & Tree, J. J. (2021). Do individual differences in face recognition ability moderate the other ethnicity effect? *The Journal of Economic Perspectives: a Journal of the American Economic Association, 47*(7), 893–907.

Coghill, R. C., McHaffie, J. G., & Yen, Y. F. (2003). Neural correlates of interindividual differences in the subjective experience of pain. *Proceedings of the National Academy of Sciences, 100*(14), 8538–8542.

Cooper, C. (2023). *An introduction to psychometrics and psychological assessment: Using, interpreting and developing tests*. Routledge.

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior face recognition ability in police super-recognisers. *Applied Cognitive Psychology, 30*(6), 827–840.

DeGutis, J., Bahierathan, K., Barahona, K., Lee, E., Evans, T. C., Shin, H. M., … Wilmer, J. B. (2023a). What is the prevalence of developmental prosopagnosia? An empirical assessment of different diagnostic cutoffs. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 161*, 51–64.

DeGutis, J., Yosef, B., Lee, E. A., Saad, E., Arizpe, J., Song, J. S., … Esterman, M. (2023b). The rise and fall of face recognition awareness across the life span. *Journal of Experimental Psychology. Human Perception and Performance, 49*(1), 22.

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. In , *44*. *Advances in experimental social psychology* (pp. 247–296) (Academic Press).

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*, Article e10629.

Fysh, M. C., & Ramon, M. (2022). Accurate but inefficient: Standard face identity matching tests fail to identify prosopagnosia. *Neuropsychologia, 165*, Article 108119.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., … Modrák, M. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808*.

Gelman, Andrew, & Pardoe, Iain (2004). *Bayesian measures of explained variance and pooling in multilevel (hierarchical) models*. Econometrics 0404002, University Library of Munich, Germany.

Grabman, J. H., & Dodson, C. S. (2024). Unskilled, underperforming, or unaware? Testing three accounts of individual differences in metacognitive monitoring. *Cognition, 242*, Article 105659.

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), Article 160923.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*, 1166–1186.
Hendel, E., & Brysbaert, M. (2024). *Towards understanding the low correlation between subjective and performance-based measures of emotion perception: Is one measure better than the other?*.
Jackson, M. C., Counter, P., & Tree, J. J. (2017). Face working memory deficits in developmental prosopagnosia: Tests of encoding limits and updating processes. *Neuropsychologia, 106*, 60–70.
Jenkins, A., Bayer, A., Tree, J., & Tales. (2015). A. Self-reported memory complaints: Implications from a longitudinal cohort with autopsies. *Neurology, 84*(23), 2384.
Jenkins, A., Tree, J. J., Thornton, I., & Tales, A. (2019). Subjective Cognitive Impairment in 55-65 year old adults is associated with negative affective symptoms, neuroticism, and poor quality of life. *Journal of Alzheimer's Disease, 67*, 1367–1378.
Kennerknecht, I., Grueter, T., Welling, B., Wentzek, S., Horst, J., Edwards, S., & Grueter, M. (2006). First report of prevalence of non-syndromic hereditary prosopagnosia (HPA). *American Journal of Medical Genetics Part A, 140*(15), 1617–1622.
Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *The Lancet Infectious Diseases, 17*(2), 161–173.
Kramer, R. S. (2023). Face matching and metacognition: Investigating individual differences and a training intervention. *PeerJ, 11*, Article e14821.
Kramer, R. S., Gous, G., Mireku, M. O., & Ward, R. (2022). Metacognition during unfamiliar face matching. *British Journal of Psychology, 113*(3), 696–717.
Kramer, R. S., McIntosh, R. D., & Nuhfer, E. B. (2024). 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, 127*(4), 2098–2108.
Kramer, R. S., & Tree, J. J. (2023). Investigating people's metacognitive insight into their own face abilities. *The Quarterly Journal of Experimental Psychology: QJEP, 77*(10), 1946–1956.
Makowski, D., Ben-Shachar, M. S., Chen, S. A., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in psychology, 10*, 2767.
McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behav Res, 52*, 2287–2305, 2020.
Megreya, A. M., & Bindemann, M. (2013). Individual differences in personality and face identification. *Journal of Cognitive Psychology, 25*(1), 30–37.
Murphy, B. A., & Lilienfeld, S. O. (2019). Are self-report cognitive empathy ratings valid proxies for cognitive empathy ability? Negligible meta-analytic relations with behavioral task performance. *Psychological Assessment, 31*(8), 1062–1072.
Murray, E., & Bate, S. (2019). Self-ratings of face recognition ability are influenced by gender but not prosopagnosia severity. *Psychological assessment, 31*(6), 828.
Murray, E., & Bate, S. (2020). Diagnosing developmental prosopagnosia: Repeat assessment using the Cambridge Face Memory Test. *Royal Society Open Science, 7*(9), Article 200884.
Nørkær, E., Gobbo, S., Roald, T., & Starrfelt, R. (2024). Disentangling Developmental Prosopagnosia: A scoping review of terms, tools and topics. Cortex; a Journal Devoted To the Study of the Nervous. *System and Behavior, 176*, 161–193.
Nørkær, E., Guðbjörnsdóttir, E., Roest, S. B., Shah, P., Gerlach, C., & Starrfelt, R. (2023). The Danish version of the 20-Item Prosopagnosia Index (PI20): Translation, validation and a link to face perception. *Brain Sciences, 13*(2), 337.
Nickels, L., Howard, D., & Best, W. (2011). On the use of different methodologies in cognitive neuropsychology: Drink deep and from several sources. *Cognitive Neuropsychology, 28*(7), 475–485.
Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of general psychology, 129*(3), 257–299.
Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic bulletin & review, 26*(2), 452–467.
Rouder, J. N., Kumar, A., & Haaf, J. M. (2023). Why many studies of individual differences with inhibition tasks may not localize correlations. *Psychonomic Bulletin & Review, 30*(6), 2049–2066.
Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science, 2*, e55.
Satchell, L., Jaeger, B., Jones, A. L., López, B., & Schild, C. (2023). Beyond reliability in first impressions research: Considering validity and the need to "mix it up with folks". *Social Psychological Bulletin, 18*, 1–21.
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), Article 140343.
Skuk, V. G., & Schweinberger, S. R. (2013). Gender differences in familiar voice identification. *Hearing research, 296*, 131–140.
Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72–101.
Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118*(3), 315.
Tagliente, S., Passarelli, M., D'Elia, V., Palmisano, A., Dunn, J. D., Masini, M., … Rivolta, D. (2023). Self-reported face recognition abilities moderately predict face-learning skills: Evidence from Italian samples. *Heliyon, 9*(3), Article e14125.
Towler, J. R., & Tree, J. J. (2018). Commonly associated face and object recognition impairments have implications for the cognitive architecture. *Cognitive Neuropsychology, 35*(1–2), 70–73.
Tree, J. J. (2011). Mental imagery in congenital prosopagnosia: A reply to Gruter, Gruter and Carbon (2011). *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 47*(4), 514–518.
Tree, J. J., & Jones, A. L. (2025). How prevalent is "other ethnicity blindness"? Exploring the extremes of recognition performance across categories of faces. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001730. Advance online publication.
Tsantani, M., Vestner, T., & Cook, R. (2021). The Twenty Item Prosopagnosia Index (PI20) provides meaningful evidence of face recognition impairment. *Royal Society open science, 8*(11), Article 202062.
Turano, M. T., & Viggiano, M. P. (2017). The relationship between face recognition ability and socioemotional functioning throughout adulthood. *Aging, Neuropsychology, and Cognition, 24*(6), 613–630.
Wingrove, J. R., & Tree, J. J. (2024). Can face recognition be selectively preserved in some cases of amnesia? A cautionary tale. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 173*, 283–295.
Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*(1), 10–65.
Woodhead, M. M., & Baddeley, A. D. (1981). Individual differences and memory for faces, pictures, and words. *Memory & Cognition, 9*, 368–370.
Young, A. W., & Burton, A. M. (2018). Are we face experts? *Trends in Cognitive Sciences, 22*(2), 100–110.
Young, A. W., Hay, D. C., & Ellis, A. W. (1985). The faces that launched a thousand slips: Everyday difficulties and errors in recognising people. *British Journal of Psychology, 76*, 495–523.