Not So Fast! Response Times in the Computerized Benton Face Recognition Test May Not Reflect Face Recognition Ability

Joseph DeGutis[1,2], Xian Li[1,2,4], Bar Yosef[1,2], Maruti V. Mishra[1,2,3]

[1]Department of Psychiatry, Harvard Medical school, Boston, MA, USA.

[2]Boston Attention and Learning Laboratory, VA Boston Healthcare, Jamaica Plain Division, 150 S Huntington Ave., Boston, MA, USA.

[3]Beyond Categories Laboratory, Department of Psychology, University of Richmond, Virginia, USA.

[4]Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD, USA.

Author Note

Address correspondence to: Joseph DeGutis,

VA Boston Healthcare System (182JP)

150 S Huntington Ave., Boston, MA, USA, 02130

E-mail: degutis@hms.harvard.edu

Abstract

Response times (RTs) are commonly used to assess cognitive abilities and have more recently been employed to assess face recognition ability. However, it is unclear whether face processing RTs predict recognition ability beyond accuracy. In the current study, we examined accuracy and RT on a widely used face matching assessment modified to collect meaningful RT data, the computerized Benton Facial Recognition Test (BFRT-c), and measured whether RTs predicted face recognition ability and developmental prosopagnosia (DP) vs. control group membership. 62 controls and 36 DPs performed the BFRT-c as well as validated measures of face recognition ability: the Cambridge Face Memory Test (CFMT) and a Famous Faces Memory Test (FFMT). We found little-to-no association between BFRT-c accuracy and RT in both controls ($r=.07$, $p=.59$) and DPs ($r=.03$, $p=.86$). In controls, BFRT-c accuracy robustly predicted CFMT performance ($r=.49$, $p<.001$), FFMT performance ($r=.43$, $p<.001$), and a CFMT-FFMT composite ($r=.54$, $p<.001$), whereas BFRT-c RT was not significantly associated with these measures (all $r's<.16$, $p's>.21$). We found that BFRT-c accuracy significantly differed between DPs and controls, but RT failed to differentiate the groups. Results in controls and DPs were unchanged after outlier removal. Further, combined scores of BFRT-c accuracy and RT (inverse efficiency score and balanced integration score) did not predict face recognition ability or DP vs. control group membership better than accuracy alone. These results suggest that the BFRT-c RT is not useful for characterizing individual differences in face recognition and, more generally, emphasizes the importance of validating RT measures before using them as individual difference/diagnostic measures.

Keywords: developmental prosopagnosia; face recognition; face perception; face matching; response time; individual differences

Not so fast! Response Times in the Computerized Benton Facial Recognition Test May Not Reflect Face Recognition Ability

Studies have shown associations between faster response times (RT) and better cognitive abilities across a variety of domains, including mathematical ability (e.g., Libertus et al., 2013; Ratcliff, Thompson, & McKoon, 2015), general intelligence (IQ; e.g. Lewis et al., 1968; Ratcliff, Schmiedek, & McKoon, 2008), numeracy tests (e.g., Gray & Reeve, 2016; Ratcliff, Thompson, & McKoon, 2015), and others (e.g., Goldhammer, 2015; Kyllonen & Zu, 2016). In the past five years, there has been increasing interest in using RT to measure both impairments (e.g., Geskin & Behrmann, 2018) and individual differences in face and object processing ability (e.g., Rossion & Michel, 2018; Meyer et al., 2021). However, there are drawbacks to using RTs such as potential speed-accuracy tradeoffs (for a review see Heitz et al., 2014) and because of the process impurity of RT relative to accuracy measures (e.g., Miller & Ulrich, 2013; Wilhelm et al., 2010). Further, in nearly all cases in the domain of face and object recognition, task RTs have not been validated as a measure of ability distinct from accuracy. The goal of the current study was to examine the computerized Benton Facial Recognition Test (BFRT-c), a version of a widely used diagnostic face matching task that emphasizes both speed and accuracy (e.g., Mishra et al., 2021; Murray et al., 2021), and to determine if RT on this task explains variance independent from accuracy in validated measures of face recognition ability. We also sought to determine if BFRT-c RT differentiates between controls and developmental prosopagnosics (DPs), individuals with lifelong face recognition difficulties (Duchaine & Nakayama, 2004).

RTs have often been used to highlight face processing deficits in prosopagnosia. In an earlier study of 5 DPs, 3 acquired prosopagnosics (APs), and 10 controls, Behrmann et al. (2005) administered both face recognition and face matching tests. Notably, these tests were given with unlimited stimulus presentation time and there were no explicit instructions to

complete each trial as fast as possible. Results showed that, in addition to having impaired accuracy on all three tests, APs and DPs were significantly slower than controls, taking on average 1.5-2.5 times longer to respond. Duchaine and Nakayama (2005) similarly found that 7 DPs were both significantly less accurate and had significantly longer RTs than controls on old/new face recognition tasks, even when instructing participants to respond as quickly as possible. Duchaine and Weidenfeld (2003) suggested that, for face matching tasks, longer RTs in prosopagnosics may be especially pronounced when a piecemeal approach to matching is possible. Using a laborious feature-by-feature approach, DPs may be able to perform better at face matching while sacrificing speed. Along similar lines, Rossion and Michel (2018) suggested that prosopagnosics, by taking their time, can perform in the normal range on the original non-speeded version of the BFRT, a widely used test of face matching (Benton & Van Allen, 1968). Indeed, normal non-speeded BFRT performance was demonstrated in 7 of 11 DPs by Duchaine and Nakayama (2004). With the goal of trying to capture prosopagnosia-related slowing and using RT as a potential diagnostic criterion, Rossion and Michel (2018) developed a computerized, speeded version of the BFRT (BFRT-c), where subjects are instructed to respond as quickly and accurately as possible. They recently published normative BFRT-c accuracy and RT[1] data (Rossion & Michel, 2018). A recent study found that accuracy on this speeded version of the BFRT-c predicted DP group membership better than other face matching tasks (Mishra et al., 2021). Further, Murray and colleagues (2021) recently found that a computerized version of the BFRT with revised, updated stimuli (BFRT-r) robustly differentiated DPs and controls. However, it remains to be seen if BFRT-c or BFRT-r RT provides unique prosopagnosia diagnostic information above and beyond accuracy and a goal of the current study is to assess this possibility.

[1] Rossion and Michel (2018) use the term "response time" to refer to the total completion time of the BFRT-c and we use the same terminology for consistency. Note that this total completion time is perfectly correlated with the more traditionally reported trial-averaged response time.

Beyond using RT to capture face processing deficits, prosopagnosia researchers have also used RT to try and determine the degree to which DPs' visual recognition deficits are specific to faces (e.g., Geskin & Behrmann, 2018; Rivolta et al., 2017). In an influential and highly debated meta-analysis, Geskin and Behrmann (2018) examined object recognition performance in 716 DP cases using both accuracy and RT results from tasks that largely emphasized accuracy rather than speed. They found that, of those studies that reported *both* accuracy and RT measures, 66.8% (159/238) of DPs exhibited object recognition deficits (either accuracy *or* RT *z*-scores < -2). However, the tasks examined were not optimized for measuring RT (they were higher difficulty and/or did not emphasize RT) and none of these studies validated RT as an independent measure of object recognition ability (i.e., demonstrated that RT explained unique variance in object recognition ability beyond accuracy). When considering object recognition accuracy impairments alone ($z < -2$) and excluding DPs with only an RT impairment, the number of DPs with object recognition impairments decreases dramatically to 22.0% (101/459). This 3-fold increase in the estimated prevalence of DP object recognition impairments when using RTs highlights the importance of determining whether task RTs provide a valid measure to classify visual recognition deficits.

In addition to discriminating between controls and prosopagnosics, RTs have also been used to characterize individual differences in face processing within the normal range of performance. For example, Wilhelm and colleagues (2010) administered a broad battery of face perception and memory tasks to two groups of healthy controls (N=151 and N=209), with some tasks focusing on accuracy and others emphasizing speed when accuracy was close to ceiling, as well as measures of object processing and general intelligence. Using structural equation modeling, they found that the best model included separate factors for face perception and memory accuracy as well as a general visual speed of processing factor.

In other words, they found that face perception accuracy reflected a face-specific ability, but face perception RT reflected more general visual processing speed. A follow-up study replicated these findings and, using drift diffusion modeling of RT from face processing speed tasks (where accuracy was close to ceiling), found that only the 'cautiousness of information processing', or where participants set their speed-accuracy tradeoff balance, was specific to faces (Meyer et al., 2019). Additional studies from this group have found that sometimes face processing task RTs reflect face-specific mechanisms (e.g., Ćepulić et al., 2018) and other times they do not (e.g., Hildebrandt et al., 2013; Rostami et al., 2017). In a recent review, Meyer et al. (2021) suggests that the recruitment of face specific mechanisms reflected in RT or accuracy depends on whether the face task is easy (≥90% accuracy) vs. difficult (≤75%). According to this model, for easy tasks face specificity is found when modeling RTs in a single face processing task but not when modeling RTs across multiple, different easy tasks. In contrast, in difficult face tasks that focus on accuracy, face specificity occurs in both single task approaches and task batteries. A goal of the current study is to test whether RTs provide information about individual differences in face recognition beyond accuracy in a face matching task that falls in between the easy and difficult cutoffs (BFRT-c control accuracy=83%, Rossion et al., 2018).

In addition to behavioral studies of RT, a recent electroencephalography study of individual differences in face processing also highlights that face task RT and accuracy may reflect different mechanisms (Dzhelyova et al., 2020). Dzhelyova and colleagues showed that a fast-periodic visual stimulation measure of face individuation (rare novel faces shown in a stream of the same repeated face identity) recorded from occipitotemporal regions was significantly correlated with BFRT-c RT, but not accuracy. This suggests that it is possible that BFRT-c RT provides distinct information to index individual differences in face recognition ability. Though this study and the studies by Wilhelm and colleagues (2010, and

others, e.g., Ćepulić et al. 2018) highlight that face processing speed and accuracy dissociate, they did not directly test whether face processing speed can explain additional variance in validated measures of face recognition ability. A study by Arizpe et al. (2018) found that RTs during different stages of the Cambridge Face Memory Test (CFMT, Duchaine & Nakayama, 2006) did indeed explain unique variance in famous face recognition above and beyond accuracy. However, the relationship observed was complex, with RT during the introduction stage negatively correlated with face recognition ability and RT during the noise stage positively correlated with face recognition ability. This, along with a recent review by Meyer and colleagues (2021), suggests that the relationship between RT and face recognition ability may be highly task dependent.

Together, the existing literature paints an inconsistent picture in which RT can at times provide useful information about face recognition abilities and prosopagnosia and other times it is unclear or possibly provides information redundant with accuracy. Beyond Arizpe et al. (2018), no studies to our knowledge have demonstrated that RT explains unique variance in face recognition ability above and beyond accuracy. While the BFRT-c is a highly sensitive face matching task (Mishra et al., 2021; as is the BFRT-r with updated stimuli, Murray et al., 2021) that was designed to produce RT data useful for characterizing individual differences in face recognition, BFRT-c RTs have yet to be validated using standard measures of face recognition. To address this issue, we administered the BFRT-c and two face recognition measures, a Famous Face Memory Test (FFMT, Mishra et al., 2019) and the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006), often thought of as the gold standard of individual differences in face recognition (Richler et al, 2015). Participants included 62 healthy controls and a relatively large group of 36 DPs. Given the established relationship between face perception (indexed by BFRT-c) and face recognition ability (indexed by CFMT and FFMT, e.g., Rezlescu et al., 2017; Schwartz &

Yovel, 2016; Young et al., 2012), the CFMT and FFMT should serve as excellent measures to validate BFRT-c RT. We examined whether BFRT-c accuracy and RT uniquely predicted DP vs. control group membership as well as individual differences in face recognition abilities in the normal population. Because BFRT-c RT may reflect independent mechanisms from accuracy (Dzhelyova et al., 2020) and because RTs for single, easier face tasks may reflect face-specific mechanisms (Meyer et al., 2021), we predicted that BFRT-c RT would explain additional variance from accuracy in individual differences in face recognition ability as well as DP vs. control group membership.

Methods

Participants

Our sample included 98 adults between the ages of 18 and 70 years old (62 Controls and 36 DPs). Individuals with DP were recruited from four sources: a) Our database of DPs who previously participated in laboratory studies, b) DPs referred to us from Dr. Matthew Peterson, who recently completed a DP study (Peterson et al., 2019), c) Individuals referred to our lab from Dr. Brad Duchaine's website, www.faceblind.org, and d) Individuals responding to our advertisement posted on public transportation. Control subjects were recruited from the community primarily through flyers and through the Harvard Decision Science Lab.

Before coming into the lab for testing, all participants underwent a pre-visit phone screening to ensure they did not meet any of the following exclusionary criteria: a history of a significant neurological disorder, lifetime moderate or severe traumatic brain injury (TBI) or mild TBI in the last 6 months, musculoskeletal or sensory impairments that would interfere with performing computer tasks, lack of English proficiency, current psychiatric disorders, diagnosed social cognitive disorders such as autism, and current dependence on alcohol or

other substances. Prior to data collection, consent was obtained for all participants according to the Declaration of Helsinki. The study was approved by the Institutional Review Board.

DP and Control Screening

Developmental prosopagnosics were screened using the 20-Item Prosopagnosia Index (PI-20; Shah, Gaule, Sowden, Bird, & Cook, 2015), Famous Faces Memory Test (FFMT), and the CFMT (Duchaine & Nakayama, 2006b). To qualify as a DP, participants had to report lifelong face recognition deficits (all but three scored > 65 on the PI-20, see Table 1), present with objective face recognition deficits on both the CFMT and FFMT ($z < -1.5$, as calculated from the control group in Duchaine et al., 2006b), and have an absence of significant neurological disorders (similar to recent studies, e.g., Stumps et al., 2020; Berger et al., 2022). All but one DP scored 44 or below (z-score < -2) on the original CFMT (Duchaine & Nakayama, 2006), indicating severe objective face recognition deficits. We also included one participant that we consider to be a mild DP who scored 46 on the CFMT, since the rest of their profile was consistent with prosopagnosia (e.g., PI-20 = 93, famous faces =.47). Removing this participant had no appreciable effects on any of the key analyses. Typically developing controls did not report any face recognition deficits in everyday life and all scored 45 or above on the CFMT. All participants had normal or corrected-to-normal vision and scored within the normal range on the Leuven Perceptual Organization Screening Test (L-POST; Torfs et al., 2014) to rule out lower-level visual causes of poor face recognition.

Testing Procedure

The experiments were implemented in PsychoPy v1.85.4 and JavaScript (for CFMT) and run on a laptop computer (34.5 x 19.5 cm display, 1920 x 1080 pixels, 60 Hz). Participants were seated 60 cm from the computer screen and instructed to indicate their responses using either a keyboard or a computer mouse. The study had three different face

recognition tests (Figure 1). Written and spoken instructions were provided. The order of the computerized tests was a) CFMT b) FFMT c) BFRT-c.

We used original CFMT (Duchaine & Nakayama, 2006) and the total score as our measure of interest. For a description of this task, please see Figure 1A. Famous face recognition was assessed using a set of 20 well-known celebrities from testmybrain.org (see Mishra et al, 2019). For each famous face presented at the center of the screen (see Figure 1B), participants were asked to make their best guess about the identity of the person by typing in the box provided and click 'submit', or to select "I don't know". For example, if the face shown was of Tom Cruise, and they could not remember the name but typed that he was the "Top Gun actor" OR "actor Cruise", they were scored as correct. After they entered a response, the correct answer/name of the person was displayed on the screen and they were prompted to click on either of the following: "I got it Right" (this was confirmed by the experimenter); "I got it wrong and I am familiar with this person"; or "I got it wrong and I am not familiar with this person". If the participant did not enter a guess but left the answer field blank and chose the option "I do not know," they were then provided with the answer on the next screen and asked to choose from either "I am familiar with this person" OR "I am not familiar with this person." We used the percent correct out of the total number of people that participants reported being familiar with.

The computerized Benton Facial Recognition Test (BFRT-c) was adapted from Rossion and Michael (2018). It was very similar to the original BFRT (Benton, 1994), except that RTs were recorded and the original instructions were changed to emphasize speed along with accuracy (see below for the on-screen instructions from the current study's BFRT-c part 2).

"You will see one face at the top of the screen that you will have to match to three faces presented below. Click on the 3 matching faces. Try to respond as quickly and accurately as possible."

The test uses grayscale photographs of unfamiliar faces (3 x 3.5 cm) presented with little visible hair and all external information cropped out. As shown in Figure 1C, each trial presents a target face at the top of the screen with six faces listed at the bottom of the screen in two rows.

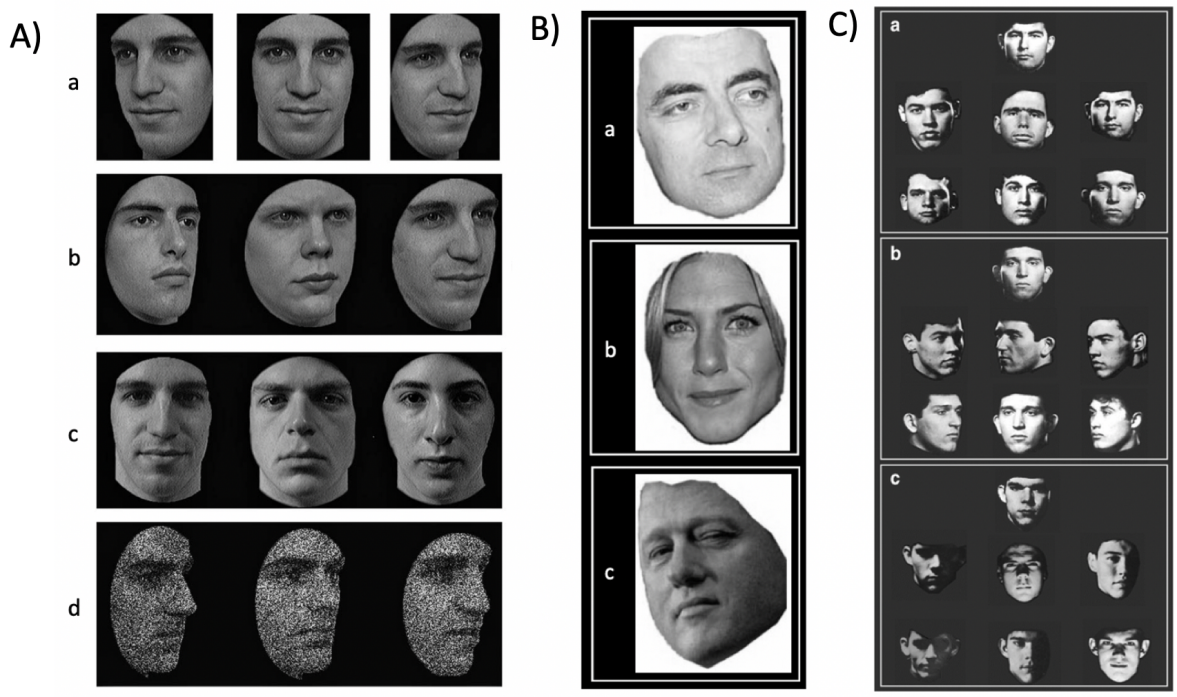


Figure 1. Example of Tasks and Stimuli. A) Cambridge Face Memory Test: a) During the learning phase, participants were shown three different images of six individuals which they were asked to memorize and b) were immediately tested on the identical image vs. foils. c) Next, participants studied the six individuals again for 20 seconds and were tested on novel images of the individuals vs. foils. d) Finally, participants studied the six individuals again for 20 seconds and were tested on novel images with visual noise vs. foils with noise. B) Famous Faces Memory Test: Participants were shown a famous face and asked to guess who this person was; after answering, the next page revealed the correct answer, and participants were asked to indicate whether they were familiar with the person. Examples: a) Rowan Atkinson, or Mr. Bean, b) Jennifer Aniston, or Rachel c) Bill Clinton, past U.S. president. C) Computerized Benton Facial Recognition Test: Participants were shown a target face at the top of the screen, and six faces at the bottom in two rows, from which participants were asked to click on the face(s) that belonged to the same person (of the target face). Examples: a) pick one out of six: one identical face of the sample person among the six options; b) pick three

out of six: three different-angled faces of the same person among the six options; c) pick three out of six: three faces of the same person in different lightings among the six options

The test consists of two sections, with the first section (6 trials) asking participant to select one out of the six faces that matches the target face (the rest of the faces had a small change in size or contrast from the target face). The second section (16 trials) asks participants to pick three out of the six options that matches the target face. In this section, the six faces had either lighting/illumination changes or head-orientation/view-point changes. The stimuli are displayed until the participant completes their response choices. Scoring was calculated by receiving either a 1 or a 0 for the first 6 trials and a score of 3, 2, 1, or 0 on the remaining trials based on how many faces were correctly identified as the target. The RTs during the second section (select three out of six) were expected to be drastically longer than RTs during the first section (select one out of six), so the total RT of all 22 trials was used instead of taking an average, similar to Rossion and Michel (2018). Note that, like Rossion and Michel (2018), we did not analyze RTs of only the correctly answered trials for two reasons: 1) During the second section, participants often take longer to make their first response than their second and third choices and because 2) Choices within a single trial are not independent from one another.

For the BFRT-c accuracy, we used the total score out of 54 items. BFRT-c RT was the total RT (6 one-response trials + 16 three-response trials) across trials. The log transformed RT was taken for the current Benton RT data (n=98) to meet the normality assumption. All statistical analyses regarding Benton RT were conducted on the natural log transformed RT, except for when explicitly stated.

Statistical Methods

To better represent an individual subject's overall face recognition ability, in addition to examining the CFMT and FFMT individually, we also calculated a composite

CFMT/FFMT score by averaging the z-scores of CFMT and FFMT. The z-scores for the CFMT and the FFMT were calculated using the mean (M) and standard deviation (SD) of the current study's control group. These three scores (CFMT, FFMT, Composite) served as our dependent measures to represent individual subject's face processing ability.

To examine whether the BFRT-c RT can reflect individual differences in face processing ability, we first used linear regressions to predict how BFRT-c accuracy and RT relate to face recognition performance measured by CFMT, FFMT, and the composite within the normal population. We report standardized beta coefficients. To examine whether BFRT-c RT contributes to differentiating DPs from controls, we first conducted independent samples t-tests between the two groups' BFRT-c accuracy and RTs. Next, we used logistic regression to assess whether BFRT-c accuracy and RT predict unique variance in DP diagnosis, or the likelihood that the participants will be categorized as a DP vs. control. McFadden's pseudo r-squared measures are reported for logistic regressions. We also report the difference in adjusted R^2 between models that include accuracy and another predictor (e.g., RT) and models that include accuracy as the only predictor.

After finding that RT did not explain unique variance in face recognition ability or DP vs. control group membership, we performed additional follow up tests to determine if combination scores of BFRT-c accuracy and RT could perform better than accuracy alone. Previously, studies have used inverse efficiency scores (IES, RT/accuracy) to examine if RTs can help explain unique or additional variance in DP vs. control performance (e.g., Rivolta et al., 2017), although other studies have suggested that IES should only be used under specific scenarios (e.g., when accuracy is very high, >.90, Bruyer & Brysbaert, 2011; Vandierendonck, 2017, 2018). Given this, we also include the Balanced Integrated Score (BIS), calculated using subtraction between the z-transformed values of RT and Accuracy (Liesefeld & Janczyk, 2019; Liesefeld et al., 2015).

Sample Size Justification

Our sample size was guided by previous studies comparing DPs and controls (Behrmann et al., 2005; Duchaine & Nakayama, 2005) as well as individual differences studies (Richler et al., 2011; DeGutis et al., 2013; Konar et al., 2010). Behrmann et al. (2005) found significant differences between the affected group (DPs and APs) and controls in face matching when using a sample of 10 in-lab controls, 5 DPs, and 3 APs, while Duchaine and Nakayama (2005) found significant DP and control differences in face recognition in a sample of 17 in-lab controls and 7 DPs. We included additional DPs in the current study to substantially increase our power to find group-level differences, since DP has shown to be such a heterogeneous group (Corrow et al., 2016). To test for individual differences associations with face recognition in controls, we used a similar-sized sample to studies that have found significant individual differences correlations with face recognition ability (N=38, Richler et al., 2011; N=43, DeGutis et al., 2013; N=48/N=77, Konar et al., 2010). Our sample size of 62 controls and 36 DPs, with alpha=.05 and power=.80, provided sensitivity to detect a medium between-groups effect size (Cohen's d=.52). Further, setting alpha=.05 and power=.80, 62 controls provided sensitivity to detect a small-to-moderate associations (Pearson's r=.25).

Results

Demographics and Diagnostic Test Performance

Participants included 36 DPs (29 female) with a mean age of 38.8 years (*SD* = 14.5), and 62 controls (37 female) with a mean age of 37.7 years (*SD* = 14.1). There were no significant differences between the two groups in age (p = .73), but there was a trend towards more females in the DP group (p = 0.06, χ^2_{Yates} = 3.62). As expected, compared to controls,

DPs performed significantly worse on the CFMT and FFMT and had more self-reported face recognition difficulties on the PI-20 (all *p's* < .001, see Table 1).

Table 1

Individual DP Scores and Mean of DP and Control Group Scores (±SD): Demographics, Self-Reported Face Recognition Abilities, Objective Face Recognition, and Face Matching Performance.

DP Sub #	Gender	Age	PI-20	FFMT	CFMT	BFRT-C	BFRT-RT	Log_e(RT)
1	F	22	88	0.27	34	39	175	5.16
2	F	29	88	0.35	37	42	226	5.42
3	F	34	75	0.33	39	37	166	5.11
4	M	61	89	0.29	38	36	363	5.89
5	F	36	93	0.54	35	35	288	5.66
6	M	33	80	0.53	36	33	358	5.88
7	M	27	80	0.19	38	47	184	5.21
8	F	46	75	0.39	34	39	197	5.28
9	F	53	86	0.40	35	39	264	5.58
10	F	26	80	0.47	42	42	221	5.40
11	F	35	81	0.45	43	47	366	5.90
12	F	30	69	0.43	41	42	199	5.29
13	F	32	58	0.56	40	38	148	5.00
14	F	27	86	0.29	44	39	243	5.50
15	F	63	63	0.20	37	38	227	5.43
16	F	31	89	0.10	37	42	219	5.39
17	F	55	96	0.47	33	34	201	5.30
18	F	39	78	0.47	33	42	245	5.50
19	F	28	80	0.27	42	47	300	5.70
20	M	37	91	0.35	33	40	152	5.03
21	F	28	80	0.47	39	42	354	5.87
22	F	64	85	0.25	43	43	273	5.61
23	F	52	87	0.12	38	41	182	5.20
24	F	25	88	0.27	44	42	136	4.91
25	M	50	82	0.20	44	39	315	5.75
26	F	33	89	0.35	39	39	191	5.25
27	F	23	92	0.38	32	35	238	5.47
28	F	70	83	0.13	36	39	479	6.17
29	F	27	76	0.08	42	38	171	5.14
30	F	38	87	0.18	44	35	211	5.35
31	F	52	92	0.40	43	34	380	5.94
32	F	64	82	0.35	39	40	319	5.76
33	M	57	87	0.47	44	48	416	6.03
34	M	20	72	0.13	40	44	227	5.42
35	F	23	93	0.47	46	39	366	5.90
36	F	25	80	0.30	44	37	255	5.54
DP (n=36)	29 F	38.8 ± 14.5	82.8 ± 8.4	.33 ± .13	39.1 ± 4.0	39.8 ± 3.8	257 ± 84	5.5 ± .3
Control (n=62)	37 F	37.7 ± 14.1	35.3 ± 7.2	.77 ± .18	59.2 ± 7.6	45.5 ± 3.7	241 ± 128	5.4 ± .4

Note: DP-developmental prosopagnosic; PI-20-Prosopagnosia Index-20; FFMT-Famous Face Memory Test; CFMT-Cambridge Face Memory Test; BFRT-c-Computerized Benton Facial Recognition Test; RT-Response time (total completion time); Log_eRT-Natural log of response time (total completion time)

Our control group mean BFRT-c score (M = 45.55, SD = 3.66) was similar to, though numerically higher than, previously reported normative data (N= 307, M = 44.81, SD = 3.44; Rossion & Michel, 2018), though the BFRT-c total RT was longer for our controls (M = 241.03 s, SD = 127.67 s) than the normative sample from Rossion and Michel (2018, N= 307, M = 180.85 s, SD = 59.87 s). This is likely because the age of the current control sample (age M = 37.73 years, range = 18–70 years, with 75% of participants above 26 years old) was significantly older than the normative data (age M = 22.62 years, range = 18–39 years, with 5% of participants above 26 years old) and age has been shown to be consistently related to face processing RT (e.g., Hildebrandt et al., 2013). Importantly, all the following key analyses of BFRT-c RT held up after controlling for age (see supplementary materials).

One of our control participants had particularly long RTs on the BFRT-c compared to the rest of the group, but their RT across the trials within the same section was evenly spread (i.e., no trial outliers), and their accuracy was 1.2 SD *above* the mean of our control group, suggesting that they had been focusing on accuracy and their data is valid. We therefore did not exclude any data points from the current analysis and took the natural log of BFRT-c RT to meet the normality assumption (see Figure 2). All statistical analyses regarding RT were performed with the natural log of BFRT-c RT except when explicitly noted.

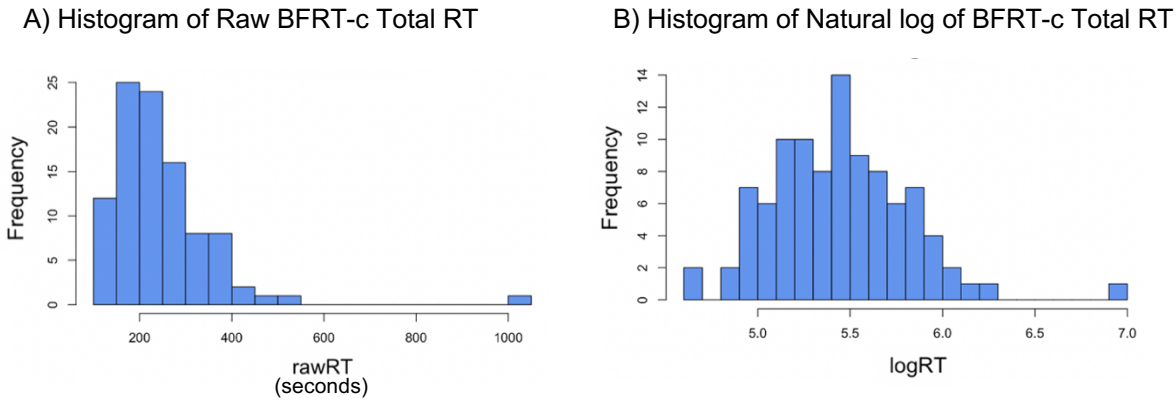


Figure 2. Computerized Benton Facial Recognition Test (BFRT-c) reaction time (RT, total completion time) distribution in the control group. A) Histogram of raw BFRT-c total RT, B) Histogram of natural log off the raw total RT.

Reliability and Association Between BFRT-c Accuracy and RT

We also examined the reliability of the BFRT-c measures using the same method as Rossion and Michel (2018), by measuring the correlation between performance on even items and odd items in the second section (pick three out of six) of the test. For BFRT-c accuracy, the interitem correlation was significant in both the control and DP groups (control: mean score even items = 19.74/24, SD = 1.99, mean score odd items = 19.85/24, SD = 2.09; $p < .001$; r_{SB} [Spearman-Brown corrected] = .63; DP: mean score even items = 17.33/24, SD = 1.90, mean score odd items = 16.72/24, SD = 2.50; $p = .016$; $r_{SB} = .57$). For BFRT-c RT, the interitem correlation was also significant, and higher, in both groups (control: mean RT even items = 12.55 s, SD = 7.75 s, mean RT odd items = 14.72 s, SD = 8.58 s; $p < .001$; $r_{SB} = .97$; DP: mean RT even items = 13.08 s, SD = 4.46 s, mean RT odd items = 14.37 s, SD = 5.19 s; $p < .001$; $r_{SB} = .95$). Note that raw RTs were used in these analyses.

Notably, we did not find any significant associations between BFRT-c accuracy and total RT in the control group (accuracy vs. RT $r = .07$, $p = .59$), DP group (accuracy vs. RT $r = .03$, $p = .86$), or for the combined Control and DP group (accuracy vs. RT $r = -.04$, $p=.70$). This suggests that there was not a significant speed-accuracy tradeoff on the task.

BFRT-c Accuracy and RT Predicting Individual Differences in Face Recognition Ability

We first sought to determine if BFRT-c accuracy and RT were associated with individual differences in objective face recognition ability in the control group, and whether RT predicted face recognition ability above and beyond accuracy. This is critical in determining whether RT carries unique information from accuracy, as researchers have suggested (Rossion & Michel, 2018). As can be seen in Figure 3, in terms of zero-order correlations, BFRT-c accuracy was significantly associated with the CFMT ($r = .49$, $p < .001$), FFMT ($r = .43$, $p < .001$), and the composite score of these two measures ($r = .54$, $p < .001$). However, the BFRT-c RT was not significantly related to CFMT ($r = -.13$, $p = .316$), FFMT ($r = .02$, $p = .889$), or the face recognition composite score ($r = -.05$, $p = .674$).

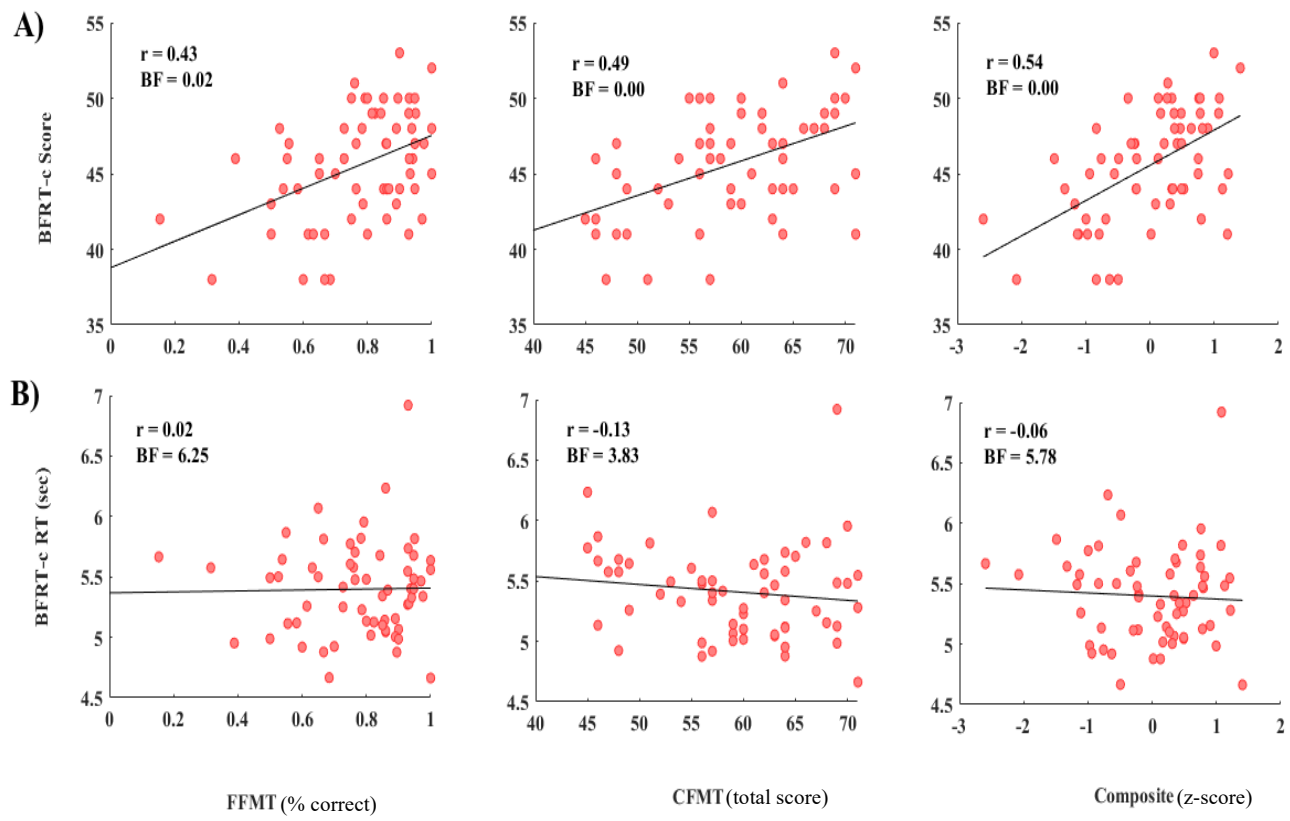


Figure 3. Associations between Computerized Benton Facial Recognition Test (BFRT-c) accuracy and RT vs. face recognition measures in the control group: A) Scatterplots for BFRT-c total score (accuracy) vs. Famous Faces Memory Test (FFMT)/Cambridge Face Memory Test (CFMT)/Composite of CFMT and FFMT ; B) Scatterplots for BFRT-c RT vs. FFMT/CFMT/Composite. BF= Bayes Factor, r = Pearson correlation.

Given that we found no significant associations between RT and any of the face processing dependent measures, to reduce the chance of falsely accepting the null (Type II error, i.e., that there was no relationship between RT and each of the dependent measure), we further computed Bayesian statistics in JASP (JASP-Team, 2018). Bayes factors showed moderate evidence in support of the null, i.e., that there was no relationship between the BFRT-c RT and the CFMT (BF_{01}= 6.25, i.e., moderate evidence for H_0), the FFMT (BF_{01}= 3.83, i.e., moderate evidence for H_0), or the Composite (BF_{01}= 5.79, i.e., moderate evidence for H_0).

When including both BFRT-c accuracy and RT as predictors of face recognition scores in a multiple regression, only accuracy was a significant predictor of CFMT (β_{acc} = .49, t = 4.36, p < .001; β_{RT} = -.13, t = -1.17, p = .248; F (2, 59) = 10.17, R^2_{acc+RT} = .26, $R^2_{acc+RT,\ adjusted}$ = .23; $\Delta R^2 = R^2_{acc+RT} - R^2_{acc}$ = .02), FFMT (β_{acc} = .43, t = 3.64, p < .001; β_{RT} = -0.01, t = -.09, p = .928; F (2, 59) = 6.65, R^2_{acc+RT} = .18, $R^2_{acc+RT,\ adjusted}$ = .16; ΔR^2 = .00), and the composite score (β_{acc} = .55, t = 5.01, p < .001; β_{RT} = -.09, t = -.83, p = .408; F (2, 59) = 12.67, R^2_{acc+RT} = .30, $R^2_{acc+RT,\ adjusted}$ = .28; ΔR^2 = .01). As shown by the change in R^2, compared to the models with only BFRT-c accuracy as a predictor, including BFRT-c RT in the model did not improve the models by a noticeable amount.

Comparing BFRT-c Accuracy and RT Between DPs and Controls

We next sought to determine whether RT would differentiate between membership in the DP (N=36) vs. control (N=62) groups and, importantly, whether RT would help

discriminate between DPs and controls above and beyond accuracy, as suggested by Rossion and Michel (2018). As can be seen in Figure 4, we found that the DP group performed much worse than the control group on the CFMT ($t_{(96)}$ = 12.8, p < .001, Mean difference, MD = 0.44, Cohen's d = 2.69), FFMT ($t_{(96}$ = 14.7, p < .001, MD = 20.10, Cohen's d = 3.08), and the composite of these measures ($t_{(96)}$ = 16.8, p < .001, MD = 2.55, Cohen's d = 3.52). While DPs also showed drastically reduced scores on the BFRT-c ($t_{(96)}$ = 7.4, p < .001, MD = 5.74, Cohen's d = 1.54), the natural log of BFRT-c RT exhibited no significant difference between DPs and the controls ($t_{(96)}$ = 1.3, p = .190, MD = .10, Cohen's d = 0.276).

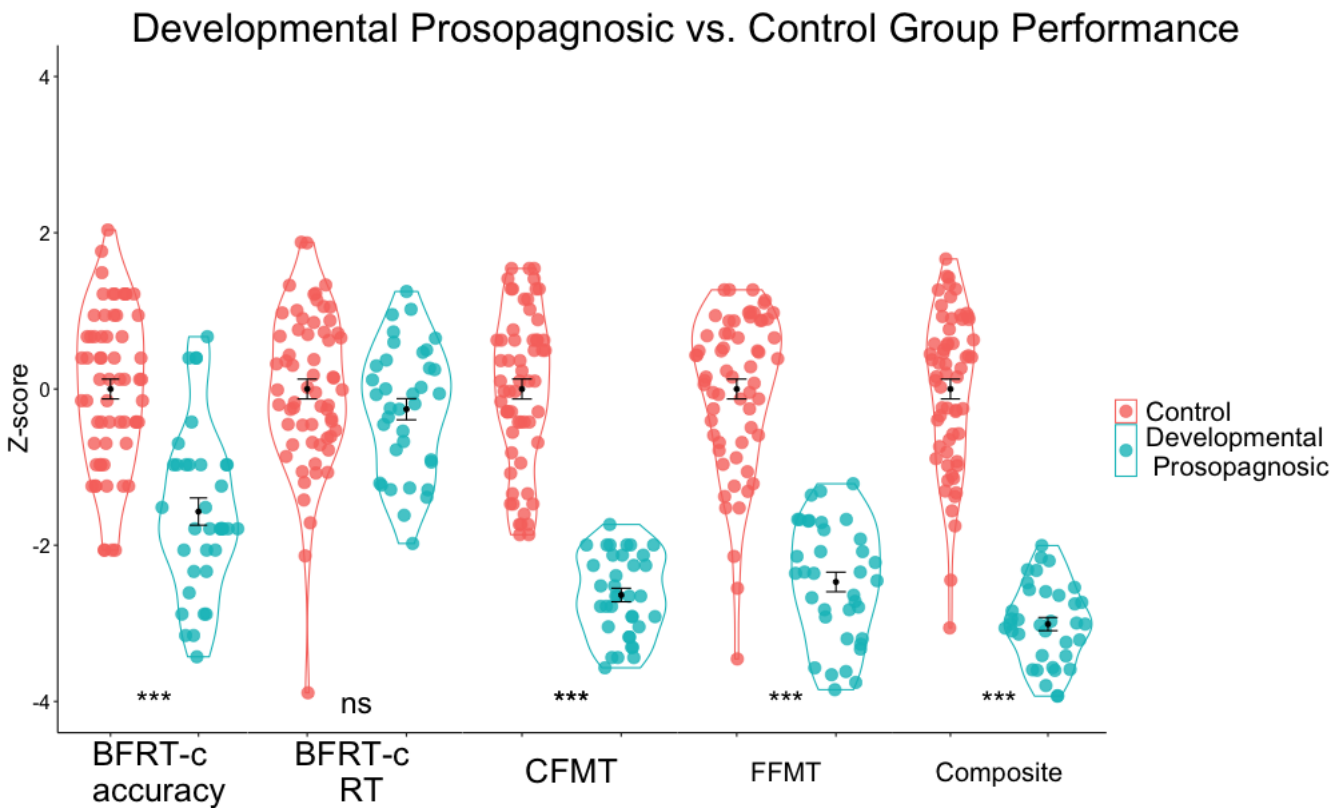


Figure 4. Control and developmental prosopagnosic z-scores (computed from the mean and standard deviation of the current control group) for the Computerized Benton Facial Recognition Test (BFRT-c) accuracy, BFRT-c RT, Cambridge Face Memory Test (CFMT), Famous Faces Memory Test (FFMT), and CFMT/FFMT Composite. The error bars represent the standard error of the mean. *** indicates that the difference is significant at p<.001

Though we did not find a significant DP vs. control BFRT-c RT difference, we did find that DPs were numerically slower. To reduce the chance of falsely accepting the null, we further employed the Bayesian independent samples t-test using JASP for the BFRT-c RT. The Bayes Factor showed anecdotal evidence in support of the null (BF_{01} = 2.13), i.e., that there was no difference between the two groups in their BFRT response time.

The binary logistic regression models predicting DP vs. control group membership showed that BFRT-c accuracy was a significant predictor ($p < .001$, $pR^2 = .32$, AIC = 91.44), while BFRT-c RT did not predict categorization as a DP/control ($p = .193$, $pR^2 = .01$, AIC = 131.14). When including both BFRT-c accuracy and RT as predictors in the model, BFRT-c accuracy, but not RT, significantly predicted DP diagnosis (accuracy: $p < .001$, RT: $p = .099$; $pR^2 = .34$, AIC = 90.69; $\Delta pR^2 = pR^2_{acc+RT} - pR^2_{acc} = .02$). As shown by the change in McFadden's pseudo-R^2 and in the AIC for model fit, compared to the logistic model that had BFRT-c accuracy as the only predictor, adding in the BFRT-c RT did not improve the model by a noticeable amount.

Alternative Methods of Examining Response Time

To further confirm the robustness of our finding of BFRT-c RT not explaining unique variance in individual differences in face recognition ability or DP vs. control group membership, we considered the BFRT-c RT data using several alternative methods. These included statistical tests on the raw, non-transformed RTs as well as removal of the RT outlier (one data point from the control group > 3SD across both groups). In addition, to test whether combined measures of accuracy and RT performed better than accuracy alone, we also examined the inverse efficiency and balanced integration scores. More details on these results can be found in the supplementary materials.

In brief, when analyzing the raw RTs as well as excluding the one outlier from the control group, we found highly similar results, where BFRT-c RT did not predict face

recognition ability above and beyond accuracy (e.g., raw RTs: CFMT $\Delta R^2 < .001$/FFMT $\Delta R^2 = .004$, *p's* > .580; outlier exclusion: CFMT $\Delta R^2 < .045$/FFMT $\Delta R^2 = .002$, *p's* >.066). When combining BFRT-c RT and accuracy into either the inverse efficiency score (IES) or balanced integration score (IES), we found that these combined measures did significantly predict face recognition ability (e.g., IES: CFMT r = -.47/FFMT r = -.33, p's < .009; BIS: CFMT r = .44/FFMT r = .30, p's < .018). However, neither measure explained unique variance beyond accuracy alone (IES: CFMT $\Delta R^2 = .02$/FFMT $\Delta R^2 < .001$, *p's* > .20; BIS: $\Delta R^2 = .02$/FFMT $\Delta R^2 < .001$, *p's* > .25).

Discussion

Despite the rising interest in using RT to examine individual differences in face processing ability and as a diagnostic measure for face and object recognition deficits, the evidence for the validity of using RT in this manner is surprisingly limited. In the current study, we examined the associations between face recognition ability with accuracy and RT from the BFRT-c, a widely used face-matching test that has been modified to create more meaningful RT data. In contrast to suggestions by the developers of this speeded, computerized version (Rossion & Michel, 2018), the current study showed no evidence that RT on the BFRT-c provides unique information about individual differences in face recognition ability above and beyond accuracy on the CFMT, FFMT, or a composite of these measures. The current study also found that BFRT-c RT did not provide any additional information beyond accuracy in diagnosing prosopagnosia, nor did the inverse efficiency or balanced integration scores (see supplementary materials), which have been recommended as an approach to incorporating RT and accuracy (e.g., Geskin & Behrmann, 2018). These results suggest that BFRT-c RT may not be a valid measure of face recognition ability and emphasize the importance of validating RT measures before using them to assess individual differences in face and object recognition abilities.

The current findings extend Meyer et al.'s (2021) model suggesting that RTs from easy face tasks (90% or greater accuracy) and accuracy from more difficult face tasks (75% or lower accuracy) are most predictive of face-specific recognition abilities. In the BFRT-c, which has an intermediate level of difficulty (M=84% for controls in the current study; M=83% for controls in Rossion et al., 2018), we found that accuracy significantly and robustly predicted individual differences in face recognition (r=.54) whereas RTs did not show any appreciable association with face recognition (r=-.06). This suggests that accuracy can be an excellent predictor of face recognition abilities in a medium-level difficulty task and further, that RTs may only be informative of face recognition abilities when accuracy is above 90%. In other words, RTs from tasks with accuracy below ceiling performance may not be particularly useful for characterizing individual differences in face recognition abilities.

The current findings suggest that recent results using RT as an indicator of face and object recognition abilities may need to be re-interpreted or at least may require additional validation (e.g., Geskin & Behrmann, 2018, Dzhelyova et al., 2020). For example, Dzhelyova et al. (2020) observed a significant association between BFRT-c RT (but not accuracy) and EEG fast periodic visual stimulation identity responses from occipito-temporal regions. The current results suggest that this association may not necessarily reflect individual differences in face recognition ability, but alternatively may reflect more general speed of visual processing abilities (Wilhelm et al., 2010). Further, Geskin and Behrmann (2018) found that when examining RT and accuracy measures in object recognition tasks, 66.8% of DPs had object recognition deficits on accuracy *or* RT (z-score < -2). This is much higher than the object deficit rate of 22% when only considering accuracy alone (z-score < -2). Notably, nearly all of the tasks included in Geskin and Behrmann's meta-analysis were designed to focus on accuracy rather than RT (e.g., accuracy was well below ceiling). The current results

emphasize the importance of evaluating each measure for whether RT can explain unique variance in face/object processing and further caution that it should not be assumed that RT will also provide unique information, even in tasks that are designed to produce meaningful RT data such as the BFRT-c.

Why did RT in the BFRT-c tell us so little about face recognition ability, while BFRT-c accuracy told us so much? It may be that instructing participants to "*Try to respond as quickly and accurately as possible"* paradoxically made the task more sensitive to accuracy and minimized RT differences in comparison to the traditional BFRT, which includes no speed-related instructions. Under the traditional BFRT instructions, it could be that to be as accurate as possible, participants would spend as much time as they needed to achieve high accuracy or high levels of confidence. This may have led to more ceiling-like effects on accuracy across the participants, which would make it less able to capture individual differences or prosopagnosia-related face matching deficits (see Duchaine & Nakayama, 2004). However, RT from the traditional BFRT (without speeded instructions) may be more meaningful than the current BFRT-c with speeded instructions, which could be empirically tested in future studies. With the instruction of speed as well as accuracy in the current BFRT-c, this may have acted as a motivational factor for the participants that minimized the differences in RT while improving accuracy in its effectiveness to reflect individual differences in facial processing ability. In support of this, the traditional BFRT has demonstrated higher average accuracies (Albonico et al., 2017; Wang et al., 2020) than the BFRT-c. Importantly, the BFRT-c accuracy did a substantially better job at distinguishing between DPs and controls in the current experiment than the traditional BFRT (Duchaine & Nakayama, 2004; see Mishra et al., 2021).

The current BFRT-c RT findings raise the question of under what circumstances should RT be used as a measure of face and object processing abilities. One reason why the

BFRT-c RT may not have explained face recognition performance is because the task is not optimized for collecting meaningful RT data, having too low of accuracy (Meyer et al., 2021) and an inability to analyze correct RTs only. To make RTs more meaningful it may be helpful for future studies to use data-limited tasks where a face is shown for a brief duration (e.g., Barragan-Jason et al., 2015). Alternatively, to better get at face-related speed of processing one could perform staircase thresholding procedures (e.g., Xu & Biederman, 2014), such as having faces be displayed for briefer durations as accurate responses are made. It would also be useful for future studies to perform drift diffusion analysis of face processing task RTs, as this has shown to reveal face-specific RT effects (e.g., Meyer et al., 2019), though this is often at the expense of requiring participants to perform many trials. Additionally, several studies from Wilhelm and colleagues (e.g., Wilhelm et al., 2010; Hildebrandt et al., 2013) have taken the approach of examining RT measures using a much easier set of face tasks with very high accuracy and separately assessing face accuracy with a harder set of perception and memory tasks. Having separate tasks may be necessary to isolate RT effects because tasks where accuracy is well-below ceiling (< 90%) may suffer from differential speed-accuracy tradeoffs across participants, limiting the usefulness of RT measures (see Draheim et al., 2019).

While integrative measures of RT and accuracy, such as IES and BIS, may be useful and can sometimes account for speed-accuracy interactions, in the current study these measures did not provide any additional information beyond accuracy in predicting prosopagnosia group membership or individual differences in face recognition in the control group. This may be due to methodological concerns with these measures. IES may only be appropriate when accuracy is very high (> 90%) and there is evidence of a speed-accuracy tradeoff (Bruyer & Brysbaert, 2011; Vandierendonck, 2017, 2018). Unfortunately, the current study met neither of these requirements. IES and BIS may have been uninformative in the

current study because BFRT-c RT and accuracy could reflect distinct cognitive mechanisms in the task (see more on this below), in which case integrating accuracy and RT in these measures may not be valid (Liesefeld & Janczyk, 2019). Further, IES and BIS are both derived measures and may have lower reliability and questionable utility in individual differences contexts (Draheim et al., 2019).

Even if a face processing task was constructed to more ideally measure RT, another reason why RT may not robustly relate to face recognition ability is that face accuracy and RT may reflect separate mechanisms and the current definitions of face recognition ability and DP group membership are based on accuracy measures (Corrow et al., 2016; Rezlescu et al., 2017; DeGutis et al., 2013). In particular, studies using structural equation modeling have shown that face RT and accuracy load on separate, largely independent factors (Wilhelm et al., 2010; Hildebrandt et al., 2013; Meyer et al., 2019), with face accuracy more consistently demonstrating face-specific loadings and RT often showing more general visual object processing loadings (though see Ćepulić et al., 2018 for an exception). This begs the question of whether face task RTs *should* be used as a measure of individual differences in face recognition ability and DP group membership. Face task RTs often provide independent information from accuracy (as in the current study as well as Dzhelyova et al., 2020), but it remains unclear if this extra information can be relevant to individual differences in face recognition. Recent work by Arizpe et al. (2018) found that overall CFMT accuracy and RTs during the 18 trials of the introduction stage of the CFMT (where accuracy is close to ceiling) explained unique variance in FFMT accuracy, though RTs explained < 5% of additional variance in FFMT. There have also been a limited number of studies that have shown RT differences between DPs and controls on face tasks with appropriately high accuracy (e.g., face detection, Garrido et al., 2008), though these effect sizes have often been much smaller than with accuracy-based tasks (e.g., Mishra et al., 2021). Thus, an important area of future

research will be to determine whether face RT tasks can reflect meaningful aspects of individual differences in face and object recognition abilities.

Though the current results provide convincing converging evidence that BRFT-c RT does not reflect face recognition abilities, there are some limitations. First, though the DP group was relatively sizable, the healthy control group could have been larger, and it would be good to replicate the current results in a larger sample of controls. Additionally, even though the current study suggests that BFRT-c RT is not explaining additional variance in face recognition, it does not mean that all face task RTs are completely uninformative. It is important for future studies to carefully examine face processing task RTs for potential outliers (e.g., similar to Dzhelyova et al., 2020) as well as potential speed/accuracy tradeoffs to better interpret the accuracy results. Finally, the current study evaluated one face task predicting face recognition ability and it would be important for future studies to examine additional face tasks along with analogous object tasks to further examine whether relationships reflect face-specific mechanisms. That being said, RTs from single face processing tasks have been shown to reflect face-specific mechanisms more than batteries of face tasks (Meyer et al., 2021), and for the current study it would be unlikely that adding an analogous object matching task to the analyses would reveal significant RT/face recognition associations.

In sum, the current results demonstrate that though the BFRT-c was modified from the original version to produce a meaningful RT measure, that BFRT-c RT failed to predict individual differences in face recognition ability or DP vs. control group membership. These results are relevant to researchers using RTs to assess visual recognition abilities and highlight the importance of testing the validity of RTs before using them as outcome measures.

Acknowledgments

We want to thank the developmental prosopagnosics and control participants for completing our challenging battery of tasks.

Declarations

Funding

This study was funded by a grant.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics Approval

The study was performed in line with the principles of the Declaration of Helsinki. The study was approved by two Institutional Review Boards.

Consent to Participate

Prior to data collection, consent was obtained for all participants according to the Declaration of Helsinki.

Consent for Publication

N/A

Availability of Data and Materials

Upon acceptance we will make the data and materials available through dryad

Code Availability

N/A

Open Practices Statement

The data and materials will be made available through dryad upon acceptance. The experiment was not preregistered.

References

Albonico, A., Malaspina, M., & Daini, R. (2017). Italian normative data and validation of two neuropsychological tests of face recognition: Benton Facial Recognition Test and Cambridge Face Memory Test. *Neurological Sciences*, 38(9), 1637-1643.

Arizpe, J., Saad, E., Wilmer, J., & DeGutis, J. (2018). The relation between facial recognition response time and facial recognition ability: task demands modulate its direction and magnitude. *Journal of Vision*, 18(10), 930-930.

Barragan-Jason, G., Cauchoix, M., & Barbeau, E. J. (2015). The neural speed of familiar face recognition. *Neuropsychologia*, *75*, 390-401.

Behrmann, M., Avidan, G., Marotta, J. J., & Kimchi, R. (2005). Detailed exploration of face-related processing in congenital prosopagnosia: 1. Behavioral findings. *Journal of cognitive neuroscience*, 17(7), 1130-1149.

Benton, A. L., & Van Allen, M. W. (1968). Impairment in facial recognition in patients with cerebral disease. *Cortex*, *4*(4), 344-IN1.

Berger, A. F. F., Fry, R., Bobak, A. K., Juliano, A., & DeGutis, J. (2022). Distinct abilities associated with matching same identity faces vs. discriminating different faces: Evidence from individual differences in prosopagnosics and controls. *Quarterly Journal of Experimental Psychology.*

Bruyer, R., & Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (IES) a better dependent variable than the mean reaction time (RT) and the percentage of errors (PE)?. *Psychologica Belgica*, 51(1), 5-13. https://doi.org/10.5334/pb-51-1-5

Ćepulić, D. B., Wilhelm, O., Sommer, W., & Hildebrandt, A. (2018). All categories are equal, but some categories are more equal than others: The psychometric structure of object and face cognition. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, *44*(8), 1254.

Corrow SL, Dalrymple KA, Barton JJ. (2016). Prosopagnosia: current perspectives. *Eye Brain*. 2016;8:165-175. Published 2016 Sep 26. doi:10.2147/EB.S92838

DeGutis, J., Wilmer, J., Mercado, R. J., & Cohan, S. J. (2013). Using regression to measure holistic face processing reveals a strong link with face recognition ability. *Cognition*, 126(1), 87-100.

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, *145*(5), 508.

Duchaine, B. C., & Weidenfeld, A. (2003). An evaluation of two commonly used tests of unfamiliar face recognition. *Neuropsychologia*, *41*(6), 713-720.

Duchaine, B. C., & Nakayama, K. (2004). Developmental prosopagnosia and the Benton Facial Recognition test. *Neurology*, 62(7), 1219-1220.

Duchaine, B., & Nakayama, K. (2005). Dissociations of face and object recognition in developmental prosopagnosia. *Journal of cognitive neuroscience*, *17*(2), 249-261.

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.

Dzhelyova, M., Schiltz, C., & Rossion, B. (2020). The relationship between the benton face recognition test and electrophysiological unfamiliar face individuation response as revealed by fast periodic stimulation. *Perception*, 49(2), 210-221.

Fry, R., Wilmer, J., Xie, I., Verfaellie, M., & DeGutis, J. (2020). Evidence for normal novel object recognition abilities in developmental prosopagnosia. *Royal Society open science*, *7*(9), 200988.

Garrido, L., Duchaine, B., & Nakayama, K. (2008). Face detection in normal and

prosopagnosic individuals. *Journal of Neuropsychology*, *2*(1), 119-140.

Garrido, L., Duchaine, B., & DeGutis, J. (2018). Association vs dissociation and setting appropriate criteria for object agnosia. *Cognitive neuropsychology*, *35*(1-2), 55-58.

Geskin, J., & Behrmann, M. (2018). Congenital prosopagnosia without object agnosia? A literature review. *Cognitive neuropsychology*, 35(1-2), 4-54.

Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: interdisciplinary research and perspectives*, 13(3-4), 133-164.

Gray, S. A., & Reeve, R. A. (2016). Number-specific and general cognitive markers of preschoolers' math ability profiles. *Journal of Experimental Child Psychology*, 147, 1-21.

Heitz, R. P. (2014). The speed–accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience, 8,* 150. http:// dx.doi.org/10.3389/fnins.2014.00150

Hildebrandt, A., Wilhelm, O., Herzmann, G., & Sommer, W. (2013). Face and object cognition across adult age. *Psychology and aging*, *28*(1), 243.

Kyllonen, P. C., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, 4(4), 14.

Lewis, M., Rausch, M., Goldberg, S., & Dodd, C. (1968). Error, response time and IQ: Sex differences in cognitive style of preschool children. *Perceptual and Motor Skills*, 26(2), 563-568.

Libertus, M. E., Feigenson, L., & Halberda, J. (2013). Is approximate number precision a stable predictor of math ability?. *Learning and individual differences*, 25, 126-133.

Liesefeld, H. R., Fu, X., & Zimmer, H. D. (2015). Fast and careless or careful and slow? Apparent holistic processing in mental rotation is explained by speed–accuracy trade-offs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41, 1140–

1151. https://doi.org/10.1037/xlm0000081

Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs (?). *Behavior Research Methods*, 51(1), 40-60. https://doi.org/10.3758/s13428-018-1076-x

Meyer, K., Schmitz, F., Wilhelm, O., & Hildebrandt, A. (2019). Perceiving faces: Too much, too fast?—face specificity in response caution. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(1), 16.

Meyer, K., Sommer, W., & Hildebrandt, A. (2021). Reflections and New Perspectives on Face Cognition as a Specific Socio-Cognitive Ability. *Journal of Intelligence*, *9*(2), 30.

Miller, J., & Ulrich, R. (2013). Mental chronometry and individual differ- ences: Modeling reliabilities and correlations of reaction time means and effect sizes. *Psychonomic Bulletin & Review, 20,* 819 – 858. http://dx.doi .org/10.3758/s13423-013-0404-5

Mishra, M.V., Likitlersuang, J., Wilmer, J.B., Cohan, S., Germine, L., DeGutis, J. (2019). Gender Differences in Familiar Face Recognition and the Influence of Sociocultural Gender Inequality. *Scientific Reports,* 9, 17884. https://doi.org/10.1038/s41598-019-54074-5

Mishra, M. V., Fry, R. M., Saad, E., Arizpe, J. M., Ohashi, Y. G. B., & DeGutis, J. M. (2021). Comparing the sensitivity of face matching assessments to detect face perception impairments. *Neuropsychologia*, *163*, 108067.

Murray, E., Bennetts, R., Tree, J., & Bate, S. (2021). An update of the Benton facial recognition test. *Behavior Research Methods*, 1-16.

Ratcliff, R., Schmiedek, F., & McKoon, G. (2008). A diffusion model explanation of the worst performance rule for reaction time and IQ. *Intelligence*, 36(1), 10-17.

Ratcliff, R., Thompson, C. A., & McKoon, G. (2015). Modeling individual differences in response time and accuracy in numeracy. *Cognition*, 137, 115-136.

Rezlescu, C., Susilo, T., Wilmer, J. B., & Caramazza, A. (2017). The inversion, part-whole, and composite effects reflect distinct perceptual mechanisms with varied relationships to face recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 43(12), 1961.

Richler, J. J., Floyd, R. J., & Gauthier, I. (2015). About-face on face recognition ability and holistic processing. *Journal of vision*, *15*(9), 15-15.

Rivolta, D., Lawson, R. P., & Palermo, R. (2017). More than just a problem with faces: altered body perception in a group of congenital prosopagnosics. *Quarterly Journal of Experimental Psychology*, *70*(2), 276-286.

Rossion, B., & Michel, C. (2018). Normative accuracy and response time data for the computerized Benton Facial Recognition Test (BFRT-c). *Behavior research methods*, 50(6), 2442-2460.

Rostami, H. N., Sommer, W., Zhou, C., Wilhelm, O., & Hildebrandt, A. (2017). Structural encoding processes contribute to individual differences in face and object cognition: Inferences from psychometric test performance and event-related brain potentials. *Cortex*, *95*, 192-210.

Schwartz, L., & Yovel, G. (2016). The roles of perceptual and conceptual information in face recognition. *Journal of Experimental Psychology: General*, 145(11), 1493.

Torfs, K., Vancleef, K., Lafosse, C., Wagemans, J., & de-Wit, L. (2014). The Leuven Perceptual Organization Screening Test (L-POST), an online test to assess mid-level visual perception. *Behavior Research Methods*, 46(2), 472–487. https://doi.org/10.3758/s13428-013-0382-6

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior research methods*, 49(2), 653-673. https://doi.org/10.3758/s13428-016-0721-5

Vandierendonck A. (2018). Further Tests of the Utility of Integrated Speed-Accuracy Measures in Task Switching. *Journal of cognition*, 1(1), 8. https://doi.org/10.5334/joc.6

Wang, P., Gauthier, I., & Cottrell, G. (2016). Are face and object recognition independent? A neurocomputational modeling exploration. *Journal of cognitive neuroscience*, *28*(4), 558-574.

Wang, L. A., Herrington, J. D., Tunç, B., & Schultz, R. T. (2020). Bayesian regression-based developmental norms for the Benton Facial Recognition Test in males and females. *Behavior research methods*, 1-12.

Wilhelm, O., Herzmann, G., Kunina, O., Danthiir, V., Schacht, A., & Sommer, W. (2010). Individual differences in perceiving and recognizing faces—One element of social cognition. *Journal of personality and social psychology*, *99*(3), 530.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., Nakayama, K., & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences of the United States of America*, 107(11), 5238–5241. https://doi.org/10.1073/pnas.0913053107

Xu, X., & Biederman, I. (2014). Neural correlates of face detection. *Cerebral Cortex*, *24*(6), 1555-1564.

Young, S. G., Hugenberg, K., Bernstein, M. J., & Sacco, D. F. (2012). Perception and motivation in face recognition: A critical review of theories of the cross-race effect. *Personality and Social Psychology Review*, 16(2), 116-142

Supplementary Materials

Results

Associations between Face Recognition and BFRT-c RT/Accuracy when Controlling for Age

When including age as a predictor of face recognition scores in a multiple regression in addition to accuracy and RT of BFRT-c, only accuracy was a significant predictor of CFMT (β_{acc} = .43, t = 3.66, p < .001; β_{RT} = -.11, t = -.94, p = .35; β_{age} = -.21, t = -1.75, p = .086; F (3, 57) = 8.04; R^2 = .26) and FFMT (β_{acc} = .33, t = 2.82, p = .0065; β_{RT} = .034, t = .30, p = .77; β_{age} = -.31, t = -2.56, p = .013; F (3, 58) = 7.03; R^2 = .23), while accuracy and age were significant predictors of the composite score (β_{acc} = .45, t = 4.165, p < .001; β_{RT} = -.047, t = -.45, p = .66; β_{age} = -.31, t = -2.75, p = .0079; F (3, 58) = 11.92; R^2 = .35). The binary logistic regression models predicting DP vs. control group membership with age included as a predictor showed that BFRT-c accuracy, but not RT or age, significantly predicted DP diagnosis (accuracy: p < .001, RT: p = .032, age: p = .066; pR^2 = .37, AIC = 88.92). Overall, controlling for age did not affect the finding that BFRT-c accuracy, but not RT predicted face recognition ability and DP diagnosis.

Examining BFRT-c Raw RT and RT with Outlier Removed

Since previous face recognition studies have only used the raw RT of the computerized Benton test, we performed the same set of analysis with the untransformed raw RT data. In the main body of the paper, we took the natural log of Benton RT to better approximate a normal distribution. Results in the control group showed that while BFRT-c accuracy significantly predicted each of the dependent measure (all ps < .001 as reported in the main body of the results section), raw Benton RT was not significantly associated with

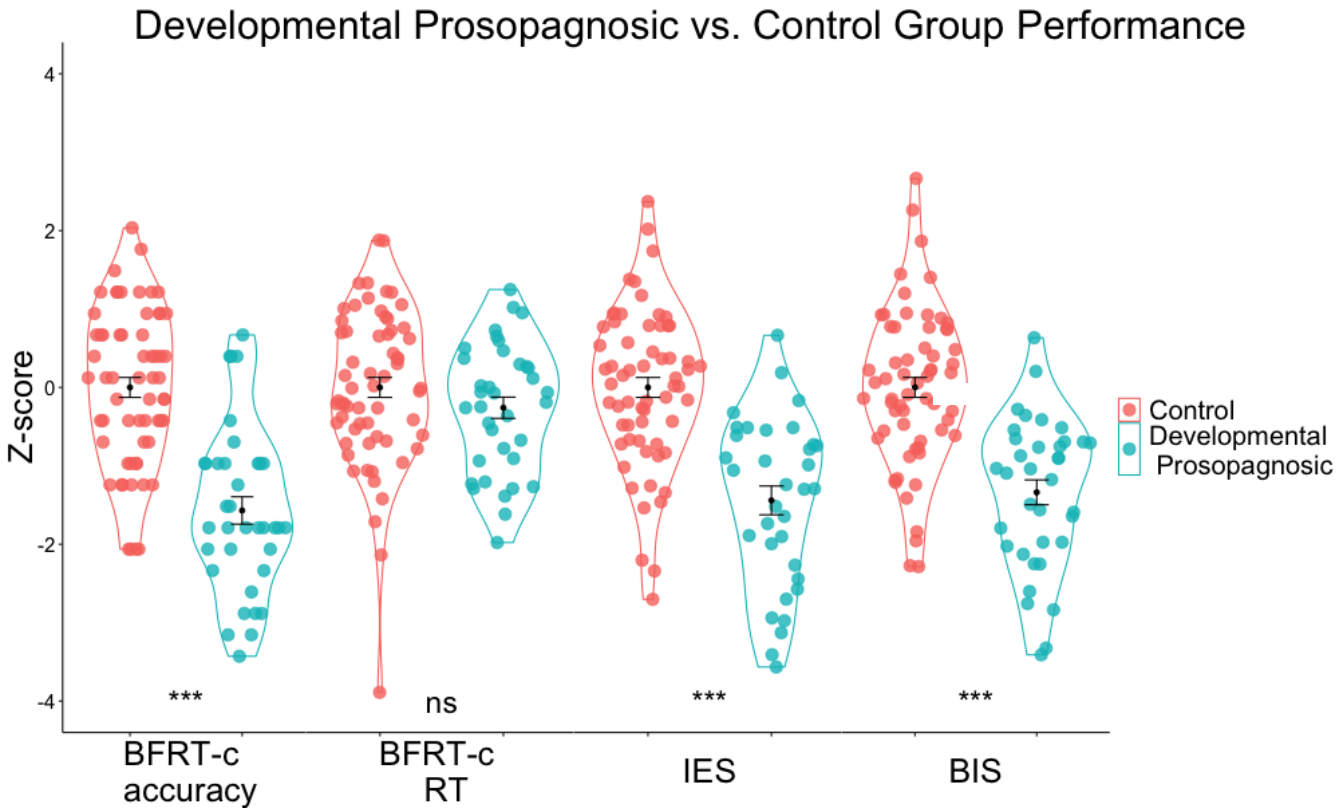
CFMT (r = -.03, p = .825), FFMT (r = .05, p = .655), or the composite CFMT-FFMT scores (r = .02, p = .863). Neither did the raw RT explain additive variance beyond accuracy for CFMT (p = .580, $\Delta R^2 < .001$), FFMT (p = .907, $\Delta R^2 = .004$), or the composite score (p = .761, $\Delta R^2 = .001$). To examine whether raw RT can differentiate DPs from the controls, we examined whether raw BFRT-c RT predicted DP/control group membership and examined whether it explained additive variance beyond accuracy in logistic regressions. Results showed that while accuracy was a significant predictor ($p < .001$ as reported in the main body of the paper) for DP diagnosis, raw BFRT response time was not a significant predictor (p = .505, $pR^2 = .003$). Neither did RT explain unique variance beyond BFRT accuracy (p = .194, $\Delta pR^2 = .011$).

In the current Benton RT data, one data point from the control group exceeded 3 standard deviations above the mean of the entire sample (i.e., combined DPs and the controls, n=98). We here provide the same set of analysis reported in the main body of the manuscript with the outlier removed RT data (natural log applied to meet normality). In the main body of the paper, we did not exclude any data based on the RT, because 1) RT was what we were aiming to validate, so exclusion criteria should not be based on this measure (exclusion is best based on the robustly validated measures); 2) excluding the outlier from only one group can artificially push the result towards one direction, and we did not want to bias our results; and 3) we examined the trials within that participant and verified that the subject was fully focused on the test, so this should be a valid data point. When examining BFRT-c RT with the outlier excluded, we found that while accuracy was significantly associated with each dependent measure (all $ps < .001$ as reported in the main body of the results section), Benton RT was only significantly associated with CFMT (r = -.26, p = .049), but not FFMT (r = -.05, p = .729), or the composite score (r = -.16, p = .215). Neither did the RT explain additive variance beyond accuracy for CFMT (p = .066, $\Delta R^2 < .045$), FFMT (p = .742, $\Delta R^2 = .002$),

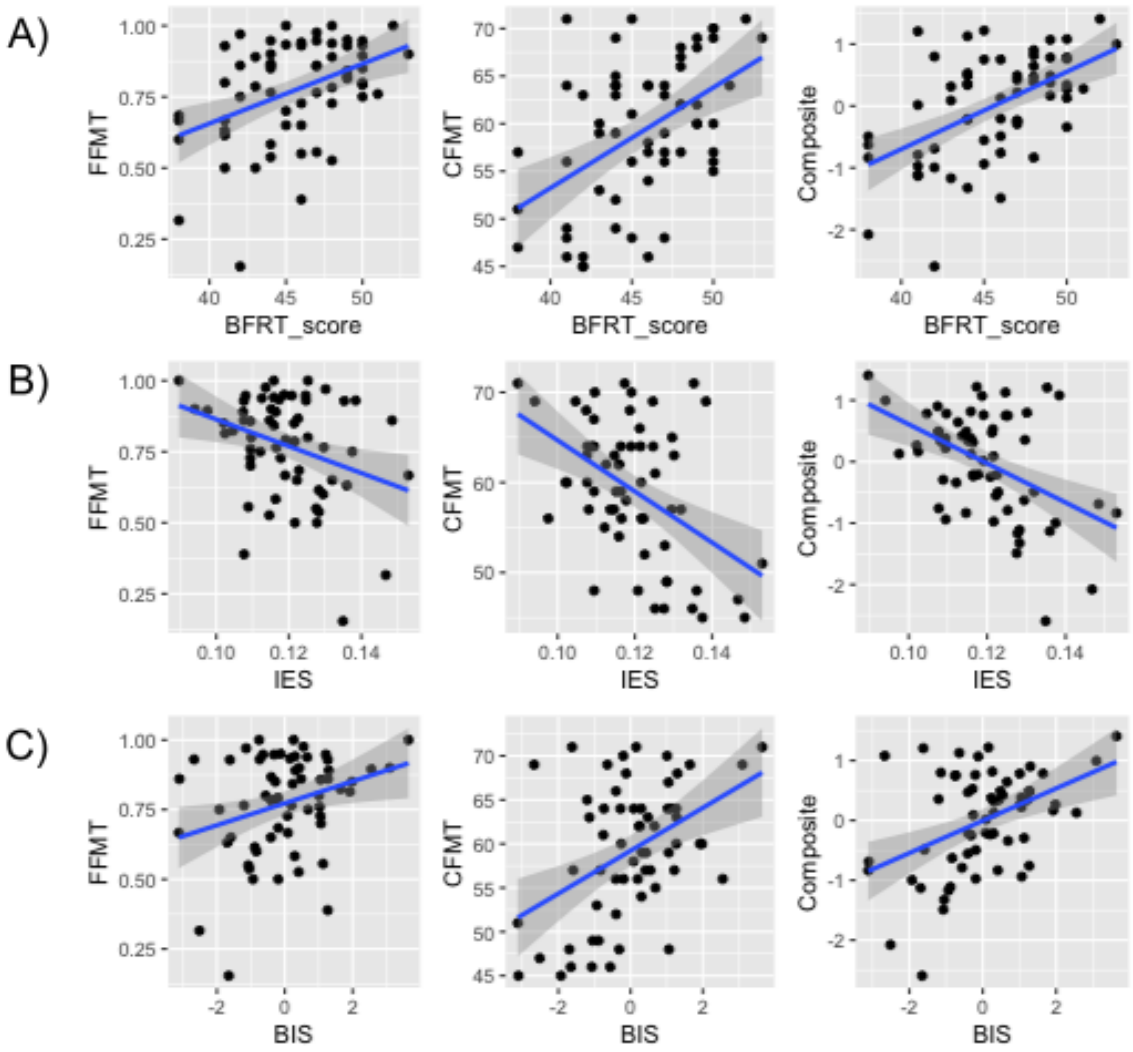
or the composite score (p = .166, ΔR^2 = .024). We next examined whether BFRT-c RT with the outlier excluded can differentiate DPs from the controls. Results showed that while accuracy was a significant predictor (p < .001 as reported in the main body of the paper) of DP vs. control group membership, BFRT-c RT was not a significant predictor (p = .077, pR^2 = .032). Neither did RT explain unique variance beyond BFRT accuracy (p = .088, ΔpR^2 = .024).

Examining Combinations of BFRT-c Accuracy and RT: Inverse Efficiency and Balanced Integration Scores

When calculating the inverse efficiency score (IES, log transformed RT/Accuracy), we found that IES did significantly predict DP vs. control group membership (p < .001, pR^2 = .302, AIC = 121.75, see Supplementary Figure S1), but adding accuracy scores in the model resulted in accuracy being the only significant predictor (accuracy: p = .0310, IES: p = .0984; pR^2 = .343, AIC = 90.623; ΔpR^2 = .02). Further, as can be seen in Supplementary Figure S2, IES was significantly correlated with FFMT (r = -.33, p = .009), CFMT (r = -.47, p < .001), and the composite of these measures (r = -.47, p < .001). However, when including both BFRT-c IES and accuracy as predictors of face recognition scores in a multiple regression, only accuracy was a significant predictor of CFMT (β_{acc} = .32, t = 1.895, p = .063; β_{IES} = -.22, t = -1.30, p = .198; F (2, 59) = 10.39, $R^2_{acc+IES}$ = .26, $R^2_{acc+IES,\ adjusted}$ = .24; ΔR^2 = .02), FFMT (β_{acc} = .41, t = 2.34, p = .0225; β_{IES} = -.03, t = -.18, p = .861; F (2, 59) = 6.67, $R^2_{acc+IES}$ = .18, $R^2_{acc+IES,\ adjusted}$ = .16; ΔR^2 = .00), and the composite score (β_{acc} = .43, t = 2.66, p = .010; β_{IES} = -.16, t = -.971, p = .335; F (2, 59) = 12.85, $R^2_{acc+IES}$ = .30, $R^2_{acc+IES,\ adjusted}$ = .28; ΔR^2 = .01). As shown by the change in R^2, compared to the model with only BFRT-c accuracy as a predictor, including BFRT-c RT in the model did not improve the model by a noticeable amount, if at all.



Supplementary Figure S1. Z-scores (computed from the mean and standard deviation of our control group) for developmental prosopagnosic and control groups on the BFRT-c accuracy, BFRT-c RT, BFRT-c IES, and BFRT-c BIS. The error bars represent the standard error of the mean. *** indicates that the difference is significant at $p<.001$



Supplementary Figure S2. Associations between face recognition measures and BFRT-c accuracy, IES, and BIS within the control group. A) Scatterplots for BFRT-c score/accuracy against FFMT/CFMT/Composite; B) Scatterplots for BFRT IES against FFMT/CFMT/Composite; C) Scatterplots for BFRT BIS against FFMT/CFMT/Composite.

We next examined the balanced integration score (BIS), calculating RT and accuracy z scores for all subjects using the Mean and SD of the control group. As shown in Supplementary Figure S1 and S2, the BIS ($p < .001$; $pR^2 = .276$, AIC = 97.254) on its own did predict DP vs. control group membership in logistic regressions and was significantly correlated with FFMT (r = .30, p = .018), CFMT (r = .44, p < .001), and the composite of these measures (r = .44, p < .001). But when accuracy was modeled along with BIS, only

accuracy was significant in predicting DP group membership (accuracy: $p = .00737$, BIS: $p = .0987$; $pR^2 = .343$, AIC = 90.689; $\Delta pR^2 = .02$). Similarly, only accuracy was a significant predictor of CFMT ($\beta_{acc} = .36$, $t = 2.274$, $p = .0267$, $\beta_{BIS} = .19$, $t = 1.167$, $p = .2481$; F (2, 59) = 10.17, $R^2_{acc+BIS} = .26$, $R^2_{acc+BIS, adjusted} = .23$; $\Delta R^2 = .02$), FFMT ($\beta_{acc} = .42$, $t = 2.601$, $p = .0117$; $\beta_{BIS} = .015$, $t = .090$, $p = .928$; F (2, 59) = 6.651, $R^2_{acc+BIS} = .18$, $R^2_{acc+BIS, adjusted} = .16$; $\Delta R^2 = .00$), and the composite score ($\beta_{acc} = .46$, $t = 3.056$, $p = .00336$; $\beta_{BIS} = .12$, $t = .834$, $p = .408$; F (2, 59) = 12.67, $R^2_{acc+BIS} = .30$, $R^2_{acc+BIS, adjusted} = .28$; $\Delta R^2 = .01$).

Together, these results suggest that, compared to accuracy alone, integrating RT and accuracy using either IES or BIS does not explain additional variance in DP vs. control diagnosis, nor does it explain additional variance in predicting face recognition scores.