New insights on real-world human face recognition

Christel Devue*, Annabelle Wride, and Gina M. Grimshaw

School of Psychology, Victoria University of Wellington

*Corresponding author: Christel Devue, PO Box 600, Wellington 6012, New Zealand. Phone: +64(0)4 463 5898. Email: christel.devue@vuw.ac.nz.

Abstract

Humans are supposedly expert in face recognition. Because of limitations in existing research paradigms, little is known about how faces become familiar in the real world, or the mechanisms that distinguish good from poor recognisers. Here, we capitalised on several unique features of the television series *Game of Thrones* to develop a highly challenging test of face recognition that is ecologically grounded and yet controls for important factors that affect familiarity. We show that familiarisation with faces and reliable person identification require much more exposure than previously suggested. Recognition is impaired by the mere passage of time and simple changes in appearance, even for faces we have seen frequently. Good recognisers are distinguished not by the number of faces they recognise, but by their ability to reject novel faces as unfamiliar. Importantly, individuals with superior recognition abilities also forget faces and are not immune to identification errors.

New insights on real-world human face recognition

Human face recognition abilities present an interesting paradox. On the one hand, we take for granted our ability to effortlessly discriminate and identify thousands of individuals with whom we are highly familiar (Ritchie et al., 2015), sometimes even decades after we last saw them (Bahrick, Bahrick, & Wittlinger, 1975). People can detect with high accuracy minute alterations to the facial configurations of famous (Ge, Luo, Nishimura, & Lee, 2003) or personally familiar faces (Brédart & Devue, 2006; Devue et al., 2007). But on the other hand, we can fail spectacularly to recognise faces of people we have just recently met, or even to match two views of the same unfamiliar individual (Bruce et al., 1999; Young & Burton, 2017; for reviews on the differences between familiar and unfamiliar face processing see Freiwald, Yovel, & Duchaine, 2016; Johnston & Edmonds, 2009; Natu & O'Toole, 2011). Processing of unfamiliar faces is easily compromised by small changes in appearance, such as the addition of glasses (Robin S S Kramer & Ritchie, 2016) or hats (as demonstrated in children; Carey & Diamond, 1977; Freire & Lee, 2001), changes in hairstyle (Toseeb, Keeble, & Bryant, 2012), lighting conditions (Johnston, Hill, & Carman, 1992), or in viewpoint (Ewbank & Andrews, 2008; Johnston, Hill, Carman, 1992).

At some point, faces must transition from fragile traces of specific encounters into robust representations of known individuals. Current hypotheses propose this process to be a computational one, whereby successive instances of a face (which may vary in viewpoint, lighting, expression, etc.) are averaged, ultimately leading to a cumulative representation based only on the invariant features that define facial identity (e.g., eyes, nose, and mouth), and enabling recognition of novel exemplars of the face (Burton, Jenkins, Hancock, & White, 2005; Burton, Bruce, & Hancock, 1999; Burton, Kramer, Ritchie, & Jenkins, 2016; Ellis, Shepherd, & Davies, 1979; Jenkins & Burton, 2011; Johnston & Edmonds, 2009). Recent studies of humans confirm that familiarisation with novel faces is facilitated by variability in views (Baker, Laurence, & Mondloch, 2017; Burton et al., 2016; Ritchie & Burton, 2016) and motion (Pilz, Thornton, & Bulthoff, 2006), and computer simulations have made good strides toward modelling the familiarisation process. However, current models do not include some of the constraints that are faced by humans (e.g., in storage capacity or degradation with time), and do not fare well with changes in peripheral features like hairstyle or facial hair that are common in the real world. Furthermore these simulations do not capture the wide range of individual differences in ability that are seen in humans. While some people - *super-recognisers* - excel at face recognition tasks (Bobak, Hancock, & Bate, 2016; Bobak, Pampoulov, & Bate, 2016; Russell, Duchaine, & Nakayama, 2009), others with developmental prosopagnosia struggle to recognise even their close relatives (Behrmann & Avidan, 2005; Susilo & Duchaine, 2013).

Better understanding of face recognition requires more information on human performance as faces transition from the novel to the familiar. However, this creates a challenge for researchers, because the process clearly requires a lot of time. Two complementary methodologies currently dominate the field. First, researchers may study recognition of highly familiar faces, often drawing on databases of celebrities. An advantage of this approach is that faces have become familiar in real-world contexts, providing ecological validity. However, conclusions are limited because this methodology provides no control over levels of exposure, nor the time course or the conditions under which faces were encountered. The alternative is laboratory-based research on recognition of novel faces. These studies provide strong experimental control and can therefore target specific perceptual or situational factors that are associated with recognition. However, they do not capture the rich context in which faces actually become familiar (Burke, Taubert, & Higman, 2007; Burton, 2013; Young & Burton, 2017), and they do not typically track changes in familiarity over the extended time course that may be required to produce robust representations. Furthermore, neither approach seriously taxes the abilities of those with superior recognition skills who tend to perform at ceiling on these tests, leaving their limits and the claim that they do not forget faces (see Russell et al., 2009) untested.

Here, we tested 'real-world' face recognition with a new task that is simultaneously ecological, tightly controlled, highly challenging (even for superior recognisers), and that assesses a range of perceptual and cognitive skills involved in recognition. We tested 32 participants who had watched the television series *Game of Thrones* (*GoT*) in its entirety, only once, as each season was released. At the time of testing, the show had run for six years and introduced over 600 previously unknown actors, who had variable (and documented) screen time. Important characters also die at alarming rates, after which the actors lose visibility. These features afford excellent control over both exposure to the faces and time elapsed since they were last encountered. Importantly, faces became familiar to viewers under naturalistic conditions (e.g., incidentally, over extended periods of time, in dynamic views, with associated changes due to aging), providing excellent ecological validity.

We presented 90 pictures of actors from *GoT* (not in character) who had four different levels of exposure in the show (main heroes, lead characters, support characters, and bit parts), and who may have appeared for the last time in any of the six seasons, mixed with 90 strangers. Participants judged whether each face was familiar, rated their confidence in that judgment, and identified and named the person if possible. To assess robustness of facial representations, half of the participants were shown pictures in which actors' headshots were similar to their appearance in the show (similar condition), while the other half saw pictures in which the actors' appearance deviated from their onscreen appearance (dissimilar condition; e.g., differences in hairstyle, facial hair, make-up, glasses, see **Figure 1**). Current research suggests that these changes might impede recognition of the least familiar faces (Clutterbuck & Johnston, 2002; Robin S S Kramer & Ritchie, 2016) but not of the most familiar faces that should benefit from robust representations. Participants also completed a commonly-used lab-based test of face recognition ability (i.e., Cambridge Face Memory Test long form - CFMT+, Russell et al., 2009), so that we could have an independent measure of their skills, and determine the relationship between performance on our ecologically-motivated task and more conventional measures of recognition ability.

Method

Participants. We tested 32 participants (20 women), aged between 19 and 56 years (Mean = 28.7 years ± 10.5), between October and December 2016, approximately 3 to 6 months after the release of the last episode of the 6th season of *GoT*. They had all watched the six seasons of the show, only once, the year each season was released (except that some watched both seasons 1 and 2 in 2012, that is the year season 2 was released, to accommodate the slow rise to popularity of the series). Moreover, they had not read George R. R. Martin's eponym books so that they would not have acquired knowledge of the characters from a different source. The recruitment notice, posted on social media and on campus at Victoria University of Wellington, mentioned a visual perception experiment but did not allude in any way to face recognition. We asked participants to come unprepared to the experiment, and not to re-watch the show or go online to study it. The sample size was constrained by the number of participants who responded to our advertisement within our testing window, and who met our strict criteria. Pilot data collected in the dissimilarity condition on 12 participants who did not meet our criteria (i.e., they had not watched the show as it was released), was noisier than our actual data would be, but revealed a non-significant association (r = .387, p = 0.214) between accuracy on our task and the CFMT+ (Russell et al., 2009). This effect size suggested that 24 or 25 participants would be sufficient to yield a significant association of similar size, allowing us to calculate correlations between our task and more conventional measures of recognition to examine individual differences. Participants provided signed consent and received course credits or movie vouchers for their time. The study was approved by the Human Ethics Committee of the School of Psychology.

Material.

Character selection and exclusion. We examined the full cast list available on the Internet Movie Database (IMDB) which details the total number of appearances of each cast member in the show (http://www.imdb.com/title/tt0944947/fullcredits?ref_=tt_ov_st_sm, retrieved September 2016) and collected more detailed information about the total amount of time that actors were visible on screen from another list available on IMDB (http://www.imdb.com/list/ls076752033, retrieved September 2016). We excluded actors who played several different characters and characters whose head was never fully visible in the show (e.g., because of headgear). Screen times were usually not available for actors who only had minor roles (i.e., “bit parts”) and were not available for all leading and support characters. We excluded potential lead characters for whom screen time was not listed.

Assignment of actors into experimental conditions (Exposure and Delay). We selected 84 actors who fit in one of 15 conditions combining 5 levels of delay since last appearance (Season 6, 5, 4, 3, 1/2) and 3 levels of exposure in the show (lead characters, support characters, and bit parts), aiming to have 5 to 6 actors per cell (see **Table 1** - the full list of actors/characters' names is presented in Supplementary material). Delay is the season in which the actor last appeared (From season 6/year 2016 to season 2 and 1 combined/years 2012-2011). Season 1 and 2 were combined in order to have enough items per cell and because many people "discovered" the show in 2012 and binge-watched the first two seasons. The three exposure bins were defined as follows. Lead characters: total screen time between 20 and 90 minutes (27 actors); support characters: total screen time between 9 and 19 minutes, or (when screen time was unavailable for 8 actors) appearing in 4 to 17 episodes (27 actors); and bit parts: appearing in one to three episodes (30 actors). Unlike extras, these latter had a role in the story and had interacted with more important characters.

In addition, we selected 6 major characters who were still alive in the last season at the time of testing (i.e., *main heroes*; screen time between 123 to 268 minutes). Because screen time accumulates from one season to another, it was impossible to manipulate delay while keeping screen time at similar levels across each season for such major characters. Main heroes were thus presented in a separate "easy" block that served to familiarise participants with the task.

Verification of actors' popularity. In order to ensure that the exposure in *GoT* matched the overall celebrity of the actor elsewhere, and especially to avoid using as bit parts actors who would be famous from other works, we used the StarMeter indicator available on IMDB. This is a ranking of actors' popularity and visibility, updated weekly, based on algorithms taking into account the number of unique viewers that visit an actor’s page, alongside other parameters (smaller ranks reflect higher popularity - e.g., lead characters’ ranks ranged from 42 to 14,373). We selected bit parts whose ranks were over 50,000 (i.e., range between 55,971 and 426,470).

A Pearson’s correlation analysis conducted on the 52 characters for whom screen times were available showed a significant negative association between StarMeter ranks and screen times in *GoT*, *r* = -.441, *p* = .001, confirming that an actor’s popularity matched their visibility in the show, and increased as their visibility in the show increased.

Similarity manipulation. We manipulated similarity in physical appearance in order to compare recognition performance for pictures showing the actor with facial features as similar as possible to their most recent appearance in the show (*similar* condition; note that just as in real life, the appearance of characters has evolved across seasons and both actors and their characters have aged) to performance with photos for which appearance was as different as possible to the character’s appearance at any time in the show (*dissimilar* condition; including variations in hair length, hair colour, hair style, presence of facial hair, glasses, apparent age, and differences in make-up that did not conceal internal features). It was impossible to manipulate similarity for bit parts because multiple pictures of these actors were not readily available.

Picture selection. The final stimulus set included 90 actors (63 men and 27 women, all Caucasian except two of mixed-race). We collected a total of 150 colour pictures of actors (i.e., not in character) from Internet sites which showed the entire face, in a frontal or slightly angled view: 12 pictures of main heroes (i.e., 6 actors x 2 pictures), 54 of lead characters (i.e., 27 actors x 2 pictures), 54 of support characters (i.e., 27 actors x 2 pictures), and 30 of bit parts (i.e., one picture per actor, which varied in similarity). We did not select pictures that served as the actor's IMDB profile picture, as it seemed likely that participants who were fans of the show might have seen this particular picture before. We also excluded pictures in which actors were in *GoT* or similar period costume, or that came from screenshots of the show (with the exception of 5 pictures for bit parts or similar versions of support characters for whom no other valid picture was available, in which case we either made sure that the costume was not visible or we altered its shape or colour slightly so that it could not be used as a recognition cue).

Table 1. Illustration of the design with Delay, Exposure, and Similarity conditions.

	Exposure			
Delay **(season last seen)**	**Main heroes**	**Lead characters**	**Support characters**	**Bit parts**
Season 6	6	6	6	6
Season 5	-	5	5	6
Season 4	-	5	5	6
Season 3	-	6	5	6
Seasons 2 & 1	-	5	6	6
Total number of actors (split by gender)	**6** (3f / 3m)	**27** (6f / 21 m)	**27** (8f / 19 m)	**30** (10f / 20m)
Median StarMeter rank	275	2,335	7,777	214,232
Similar and dissimilar pictures available	Yes	Yes	Yes	No
Total number of actors' pictures	**12**	**54**	**54**	**30**

Note. Cells show number of suitable characters selected per condition. The last four rows indicate, from top to bottom, the total number of actors, median StarMeter ranks (smaller ranks indicate higher popularity), whether similar and dissimilar pictures were available, and total number of pictures used per exposure level.

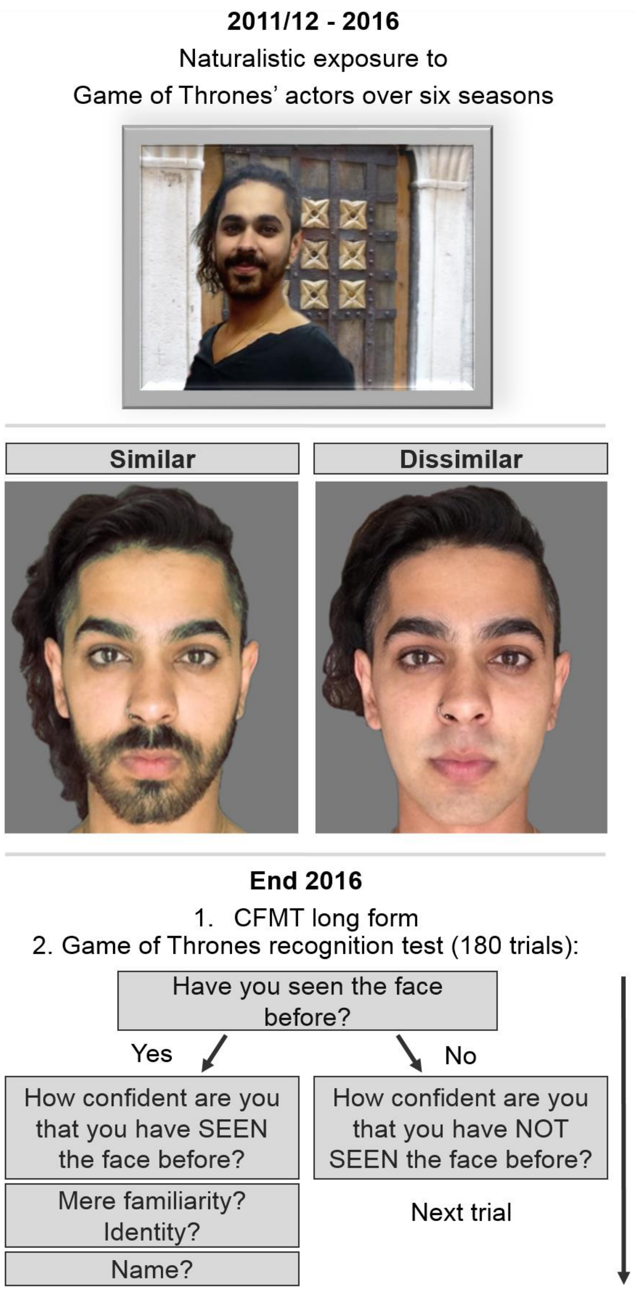

2011/12 - 2016
Naturalistic exposure to
Game of Thrones' actors over six seasons
Similar
Dissimilar
End 2016
1. CFMT long form
2. Game of Thrones recognition test (180 trials):
Have you seen the face before?
Yes
No
How confident are you that you have SEEN the face before?
How confident are you that you have NOT SEEN the face before?
Mere familiarity? Identity?
Name?
Next trial

Figure 1. Illustration of the procedure and stimuli. Top panel shows what an actor might look like in the show (mock screenshot). Middle panels represent stimuli presented in each similarity condition, where the actors' headshots either matched the appearance they had in the show (similar condition) or deviated from their look in the show (dissimilar condition - here, change in facial hair). Due to copyright restrictions, pictures used in the experiment are not shown, but they are available from the corresponding author. The model depicted here has provided permission. Bottom panel shows the testing phases and the different aspects assessed during each trial of the Game of Thrones test (grey boxes).

In addition, we collected 90 pictures of unfamiliar faces (65 men and 25 women, all Caucasian) from various websites outside New Zealand (e.g., hairstyling, amateur photograph or modelling, city councils and CEO board members pages), and strived to match the set of unfamiliar faces with the set of actors in terms of head orientation, age range, facial expression, attractiveness, presence of make-up, facial hair, or glasses, hairstyle, clothing style, lighting, and picture quality.

Picture stimuli preparation. The set of 240 photographs was then standardised. Images were rotated so that the eyes were aligned on a horizontal axis and background was replaced by a uniform mid-grey field, see middle panels of **Figure 1**. Pictures were cropped (above the head, so that hairstyle was visible but in a way that minimised the amount of visible clothing, and right under the chin or keeping a bit of the neck for people with longer hair) within a frame with a 2.5 by 3 ratio and resized to 200 x 240 pixels (i.e., about 6.5 by 7.8 degrees at 70 cm). An overview of the stimulus set is available here: https://osf.io/xjmzp/

Cambridge Face Memory Test long form (CFMT+). This test assesses recognition of novel faces (Russell et al., 2009) and was used as an independent measure of face recognition abilities. Participants study 6 male faces in 3 different viewpoints. Recognition is tested across 102 forced-choice trials, in which one of the 6 studied faces appears amongst 2 foils. Over the trials, test pictures show increasing levels of changes (e.g., in lighting, facial expression, head orientation, inclusion of hair, external features cropped out) and/or degradation (e.g., addition of digital noise).

Procedure. Participants were tested individually in a dimly lit room on a PC and sat at approximately 70 cm from the screen (resolution 1024 x 768). They first performed the upright version of the CFMT+. Then, they were assigned to one of two similarity conditions (*similar* or *dissimilar* photos) in the *GoT* test, in such a way that scores on the CFMT+, as well as age and gender, were similarly distributed in both groups (similar condition: *N* = 16, 6 men, Mean age = 29.5 years ± 9.6; dissimilar condition: *N* = 16, 6 men, Mean age = 27.9 years ± 11.5).

The *Game of Thrones* test began with an easy block in which the six main heroes were intermixed with 6 strangers (i.e., unfamiliar faces) to familiarise participants with the task. The remaining 84 characters were intermixed with 84 strangers. We created two different pseudo-randomised lists, counterbalanced across participants, in which a maximum of 4 actors or strangers were presented consecutively.

Each trial started with a fixation cross (500 ms), followed by a picture stimulus in the centre of the screen until the participant's response or up to 3000 ms, after which participants were prompted to respond. Participants pressed "K" if they had seen the face before (in *GoT* or elsewhere), and "L" if they had not. Then, they rated their confidence in this familiarity judgment on a 5-point Likert scale (1 = not at all confident that they have seen/not seen the face before, 5 = totally confident that they have seen/not seen the face before). If the face was judged familiar, they explained the nature of their recognition: mere familiarity or identity. Since faces belonged to a closed set, we requested specific semantic information (other than the name) to accept a response as an identification (e.g., one of the two characters from the Night's Watch who was always on Jon Snow's side). Responses placing the actor in a broad context of the show without individualising information (e.g., someone in the Night's Watch) were categorised as mere familiarity. To discourage participants from basing their familiarity judgment on their ability to place an actor in *GoT* (i.e., involving source memory), they were told that semantic information about the actor from outside the show was acceptable. When participants did not spontaneously produce the name in their response, they were prompted to do so.

After the *GoT* test, we also collected data from other tasks (CFMT+ inverted, CFPT, CCMT) but they are not relevant to the question addressed here, and are not discussed further.

Measures and analyses.

Recognition: Familiarity, identification and naming. Since we did not have precise a priori hypotheses on the way exposure and similarity would affect the category of responses, we conducted descriptive analyses aimed at assessing the effect of exposure to a face on people's ability to recognise it as familiar, to further identify it, and to name it. In order to have more data per exposure condition and for the sake of clarity, we did not include delay in this analysis. To preserve control over exposure, we discarded any trial in which participants subsequently reported only semantic information about a person from outside *GoT* (i.e., 26 cases across participants; 0.45% of total trials). We present mean proportions of each recognition type, calculated against the total number of usable trials presenting actors in a given condition, see **Figure 2**.

"Seen/familiar" responses were either correct, i.e., *hits* - for actors, or incorrect, i.e., *false alarms* - for strangers. Correct familiarity judgments of actors could reflect a *feeling of mere familiarity,* or could be accompanied by the ability to provide semantic information about the character but not their name, i.e., *correct ID* on **Figure 2**.

We checked the accuracy of information provided about a *GoT* character by means of a wiki website devoted to the show (http://gameofthrones.wikia.com/wiki/Category:Characters).

Correct familiarity judgments could also be accompanied by incorrect identification when the information provided did not match the character, i.e., *confusion*.

Further, participants were sometimes able to identify the person *and* provide their name, i.e., *correct ID & name* on **Figure 2**. Because some characters' names are foreign or unconventional, and may have not been encountered in a written form, we accepted the following responses as correct: a correct character's first name by itself because first names are individuating in the show and because some characters do not have a last name (e.g., 'Davos' for 'Ser Davos Seaworth'), any other name the character officially goes by that appears on the *GoT* wiki page (e.g., 'the King-Beyond-the-Wall' for 'Mance Rayder'), a first name accompanied by the wrong last name providing that it makes sense in the context of the show (e.g., 'Catelyn Tully' which is the maiden name of 'Catelyn Stark'), and names with small phonetic variations as long as the name maintained the same root (e.g., 'Oberon' instead of 'Oberyn'). Some bit parts (N = 14) were not named in the show and their character was casted under an individual label (e.g., 'Braavosi Captain), that we also accepted as a correct name. For all actors, we also accepted their full correct name (i.e., first name and last name). We did not accept a character's last name by itself, even if correct, because it is not individuating (e.g., 'the Lannister patriarch' for 'Tywin Lannister'), or names with more substantial variations (e.g., 'Lea' instead of 'Shea').

We opted *not* to calculate identification and naming performance contingent on the proportion of hits. Because the task is so challenging, some participants might have very low hit rates, but nonetheless be able to correctly identify and name all of the few faces they recognise, which would give a misleading impression of their ability (e.g., identifying 50% of recognised faces reflects different abilities depending on whether 2 or 60 out of 90 faces are recognised).

Finally, incorrect familiarity judgments for strangers could also reflect a *feeling of mere familiarity* or an incorrect identification, i.e., *intrusion*.

Sensitivity (d') and Criterion. We calculated *d'* on the basis of hit rates (i.e., correct recognition of actors) in each of the four exposure conditions, and on the basis of false alarm rates (i.e., incorrect "familiar" responses to strangers) separately in the easy block (i.e., showing 6 main heroes and 6 strangers) and in the remaining series of trials (i.e., presenting 84 actors and 84 strangers). To optimise control of exposure, we excluded trials on which faces were subsequently identified only outside *GoT*.

Because we did not have access to two photos and cumulative screen times for all characters (i.e., bit parts, or main heroes, respectively), we could not create a full factorial design crossing all the levels of our three variables of interest (i.e., exposure, delay, and similarity). Therefore, we performed two sets of analyses to examine the effect of exposure and of delay on sensitivity (*d'*) separately. First, we conducted an Analysis of Variance (ANOVA) with exposure (4 levels: main heroes, lead characters, support characters, and bit parts)[i] as within-subject factor and similarity as between-subject factor, see **Figure 3** (top panels). Second, we conducted an ANOVA with delay (5 levels: seasons 6, 5, 4, 3, 2/1) as within-subjects factor and similarity as between-subjects factor, only including trials showing lead and support characters, see **Figure 3** (bottom panels).

[i] For bit parts, since participants in the similar and dissimilar conditions saw the same photos, differences in performance will reflect the effect of the participant's condition (similar/dissimilar), and not differences in the photos of bit parts themselves.

We calculated Criterion *c* based on mean proportions of hits and false alarms across all the conditions, excluding the first easy block of trials (i.e., presenting 6 main heroes and 6 strangers). We compared criterion *c* in the two similarity groups with a Student t-test.

Confidence ratings. For each participant, we calculated mean confidence ratings for correct and incorrect familiarity judgments. Confidence ratings given after correct and incorrect responses were compared with a Wilcoxon test. Further, confidence ratings in the two similarity conditions were compared with Mann-Whitney U, see **Table 2**.

Individual differences. The goal of the following analyses was to examine which cognitive components of the task involved in person identification are associated with accurate recognition within the *GoT* test, and whether and how scores on the standard test (CFMT+) reflect performance on our more ecological test. In order to reduce the number of analyses and decrease the likelihood of Type I errors, we did not take delay or exposure into account but looked at overall performance. Moreover, because cases of misidentification (i.e., confusions and intrusions) were limited in number, all measures used in correlational analyses included all trials, even if actors had only been identified (or misidentified) outside *GoT*. Student t-tests presented in **Table 2** show which measures were affected by similarity. To maximise power, we collapsed responses from the two similarity groups and calculated Z-scores for each measure, taking the means and standard deviations of each similarity group into account.

We used Spearman's correlation analyses to examine associations between standardised accuracy in familiarity judgments/CMFT+ scores and standardised d', proportion of hits, proportion of false alarms, Criterion *c*, proportion of reported mere familiarity (for actors and strangers separately), proportion of familiar faces correctly identified (with or without naming collapsed), proportion of familiar faces correctly named, proportion of confusions, proportion of intrusions, mean confidence for accurate familiarity judgments, and mean confidence for incorrect familiarity judgments.

Results

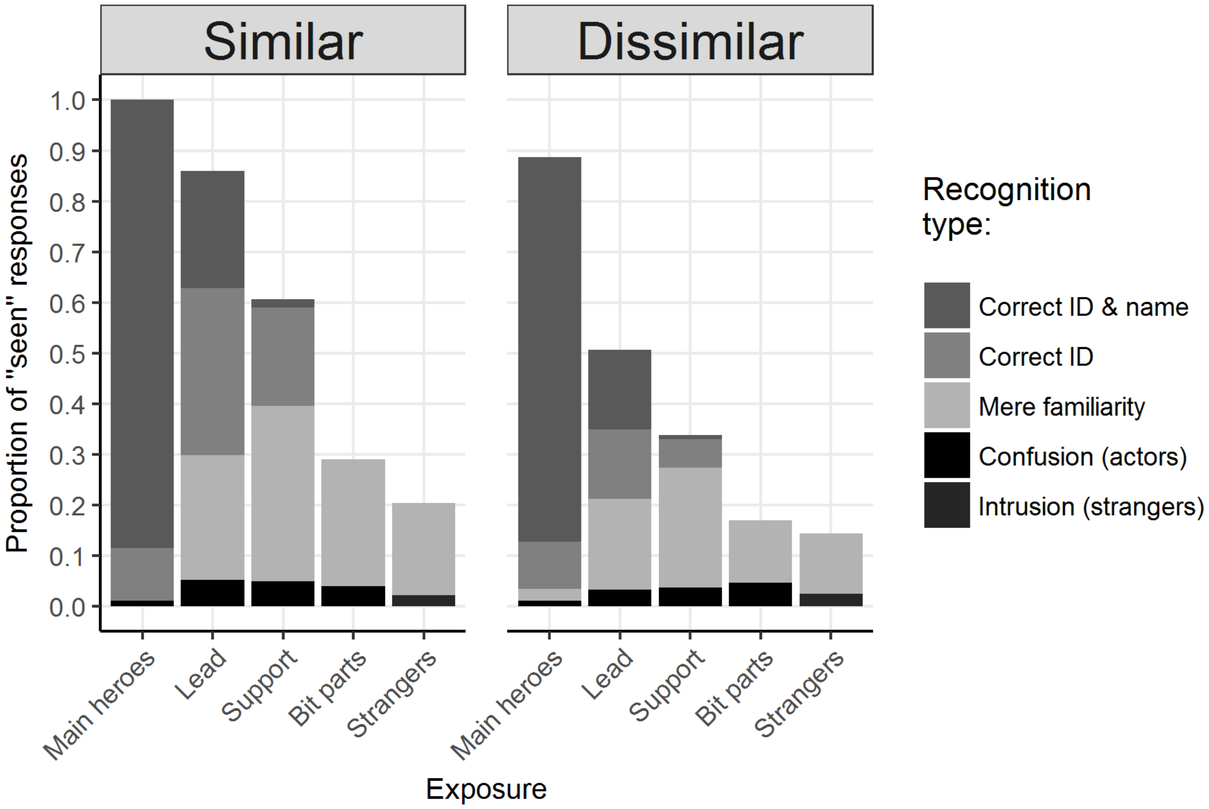


Figure 2. Results of the familiarity judgment, i.e., proportions of “seen” responses, split by recognition type, in each picture similarity condition. The first four bars show responses to actors with different levels of exposure (the height of each bar corresponds to the total proportion of hits) and the rightmost bar are responses to unfamiliar faces (so the top of the bar indicates total proportions of false alarms). People reported feelings of mere familiarity when they could not place the person. ‘Correct ID’ refers to correct identifications of an actor’s *GoT* character. ‘Confusions’ are identifications that do not match the actor. ‘Intrusions’ are erroneous identifications of unfamiliar faces.

Recognition: Familiarity, identification and naming. Descriptive statistics presented on **Figure 2** and in **Table 2** indicate that it takes extensive exposure for a face to be distinguished from novel faces and correctly identified. The least exposed actors (bit parts) were never correctly identified. Further, overall bit parts evoked feelings of familiarity (*Mean* = 18.7%, *SD* = 14.1) in the ballpark of strangers (*Mean* = 15%, *SD* = 12.1), although slightly more so, $t(31) = -2.04$, $p = .05$, $d = .36$. Main heroes and lead characters were the only classes to be correctly identified and named more often than they were merely recognised. Identification errors (confusions for actors, or intrusions for strangers) were found at all exposure levels.

Sensitivity (d') and Criterion. *Exposure.* The ability to discriminate actors from strangers as indexed by *d'* increases as a function of exposure, $F(2.172,65.173\ ) = 148.97$, $p < .001$, $\eta_p^2 = .832$, in a linear fashion, $F(1,30) = 259.74$, $p < .001$, $\eta_p^2 = .896$, but unexpectedly, changes in appearance disturb recognition across the board: more similar pictures of actors are always better discriminated from strangers than dissimilar ones, $F(1,30) = 8.15$, $p = .008$, $\eta_p^2 = .214$, see **Table 2** and **Figure 3** (top panels), and similarity does not interact with exposure, $F(2.172, 65.173) = 1.179$, $p = .317$, $\eta_p^2 = .038$.

Criterion. Note that photos of actors who played bit parts are the same in both the similar and dissimilar conditions (due to their having fewer photos available), and yet they also show worse recognition in the dissimilar condition. This might be due to changes in decision criteria depending on the task context, see Criterion *c* in **Table 2**. Where recognition is harder (i.e., dissimilar pictures), people become more cautious in their familiarity judgments than in conditions that allow for better recognition (i.e., similar pictures), $t(30) = 3.184$, $p = .003$, $d = 1.456$.

Table 2. Comparison of performance on the *Game of Thrones* test in the two picture similarity groups, and associations between different measures on the Game of Thrones test and accuracy/CFMT+ scores.

	Similar	**Dissimilar**	***Effect of similarity***	***r Accuracy***	***r* CFMT+**
Familiarity judgment					
Accuracy	0.70 ± *0.05*	0.61 ± *0.05*	**5.273*****	-	**.456****
Sensitivity (*d'*)	1.25 ± *0.34*	0.85 ± *0.45*	**2.666***	**.832*****	**.485****
Hit rate	0.60 ± *0.10*	0.37 ± *0.10*	**6.537*****	.255	-.081
False alarm rate	0.20 ± *0.15*	0.14 ± *0.09*	1.231	**-.557*****	**-.467****
Criterion (*c*)	0.35 ± *0.39*	0.78 ± *0.36*	**-3.184****	.259	.335†
Confidence accurate	3.73 ± *0.36*	3.74 ± *0.44*	128	.125	.339†
Confidence inaccurate	3.23 ± *0.40*	3. 47 ± *0.14*	111	-.096	.085
Recognition types					
Actors					
Mere Familiarity	0.26 ± *0.11*	0.17 ± *0.08*	**2.663***	-.233	**-.434***
Correct identification	0.30 ± *0.08*	0.17 ± *0.08*	**4.73*****	**.693*****	**.418***
Naming	0.13 ± *0.05*	0.10 ± *0.05*	1.916†	**.574*****	.208
Confusion	0.04 ± *0.03*	0.04 ± *0.03*	0.722	-.143	.034
Strangers					
Mere Familiarity	0.18 ± *0.14*	0.12 ± *0.08*	1.446	**-.555*****	**-.481****
Intrusions	0.02 ± *0.02*	0.03 ± *0.03*	-0.644	-.219	-.042
CFMT+ score (%)	70.65 ± *11.31*	70.47 ± *11.02*	.047	-	-
Scores range	50.98 - 87.25	52.94 - 88.24			

Note. The first two columns report performance on the recognition task in each similarity group: mean accuracy, *d'*, proportions of actors correctly recognised (hit rate), proportions of strangers erroneously recognised (false alarm rate), Criterion *c*, confidence about correct and incorrect familiarity judgments (1 = low; 5 = high); recognition category for actors and strangers (expressed as proportions out of 90 items); and CFMT+ scores (in percentage). Standard deviations are in italics. The effect of similarity is tested with Student t-tests for independent samples (2-tailed, df = 30), except confidence ratings, tested with Mann-Whitney U. The two rightmost columns show Pearson's correlation coefficients (*r*) testing associations between standardised performance (Z-scores) on the *GoT* test and overall accuracy, and between performance on the *GoT* test and CFMT+ scores, respectively. †$p < .1$, *$p < .05$, **$p < .01$, ***$p \leq .001$.

Delay. Again, there was a main effect of Similarity whereby dissimilar pictures of actors were more difficult to discriminate from strangers than similar ones, $F(1,30) = 14.938$, $p = .001$, η_p^2

= .332. A main effect of delay suggests that once exposure to faces has ceased, they tend to be forgotten, *F(*4,120) = 13.334, *p* < .001, η_p^2 = .308. The bottom panels of **Figure 3** show a recency effect whereby there is a linear decrease in recognition across seasons, *F*(1,30) = 53.84, *p* < .001, η_p^2 = .642. This linear effect interacts with similarity, *F*(1,30) = 10.05, *p* = .003, η_p^2 = .251. However, **Figure 3** suggests that the interaction is driven by unexpectedly low recognition of actors with dissimilar appearance last seen in season 5.

Individual differences. We tested associations between face recognition skills as measured with the current benchmark test, the CFMT+ (Russell et al., 2009), and accuracy on our own test, see **Table 2** and **Figure S1**.

CFMT+ scores. Good recognisers (as classified by the CFMT+) provide more accurate familiarity judgments in the *GoT* test, *r* = .456, *p* = .009, and are better at discriminating actors from strangers (see *d'*) than people with lower scores, *r* = .485, *p* = .005. Remarkably, these associations are driven by false recognitions of strangers (i.e., false alarms), *r* = -.467, *p* = .007, which are more likely in poor than better recognisers, rather than by correct recognition of actors (i.e., hits), *r* = -.081, *p* = .661, see **Figure 4**. The better predictive power of false alarms compared to hits is confirmed by a Steiger's test comparing the strengths of the associations between CFMT+ scores, and hit rates and false alarm rates, respectively (while accounting for the correlation between these two, *r* = .633, *p* < .001), *Z* = 2.635, *p* = .0084.

Furthermore, better recognisers are less likely to report feelings of mere familiarity for both actors, *r* = -.434, *p* = .013, and for unfamiliar faces, *r* = -.481, *p* = .005, than poor recognisers; and are more likely to provide correct identifying information for actors, *r* = .418, *p* = .017. By contrast, CFMT+ scores do not significantly predict one's ability to name familiar faces, *r* = .208, *p* = .253, nor, importantly, the occurrence of identification errors for actors, i.e., confusions, *r* = .034, *p* = .853, or strangers, i.e., intrusions, *r* = -.042, *p* = .819, see **Table 2** and **Figure 5**.

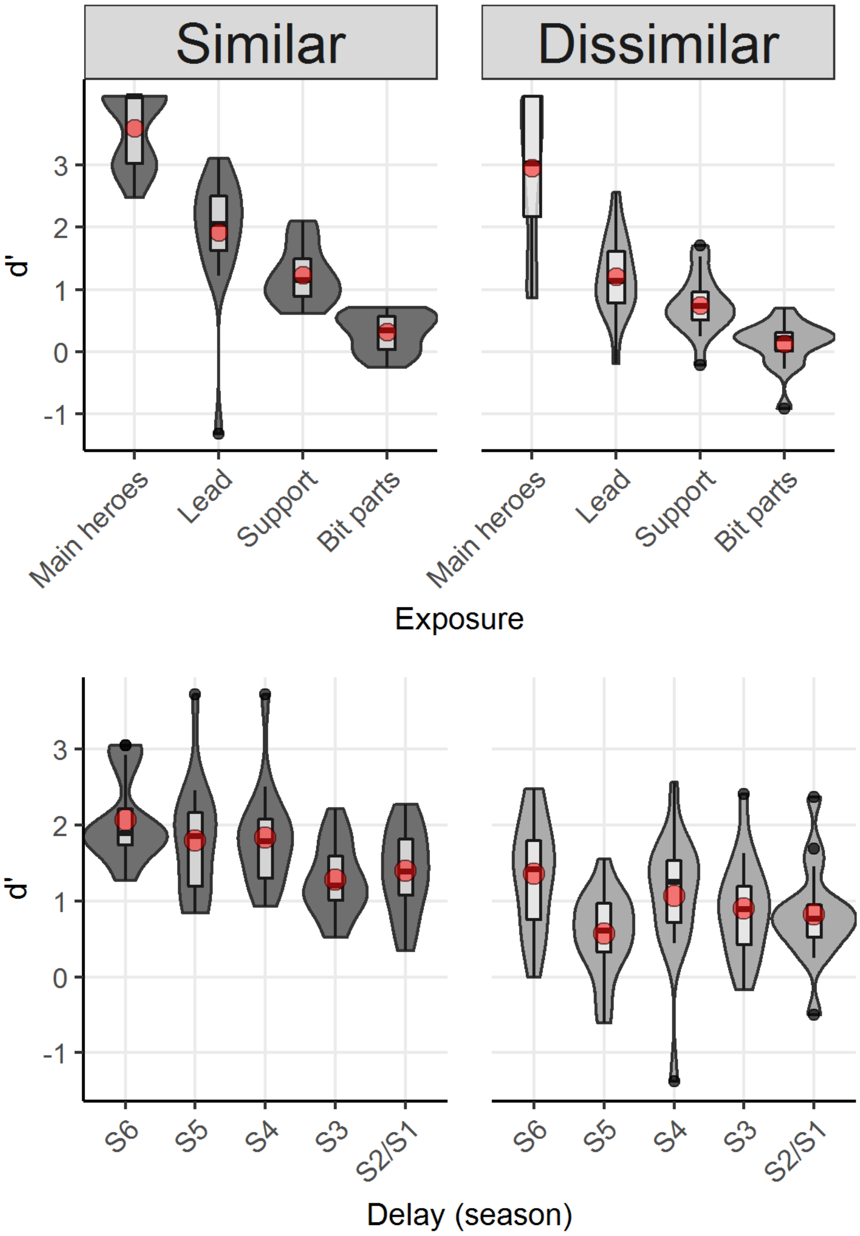


Figure 3. Discrimination performance (*d'*) as a function of exposure (top) and delay (i.e., season in which a character was last seen; bottom) in each similarity condition. Note that the latter analysis excludes main heroes who had accumulated the highest screen times across 6 seasons, and bit parts. Red circles represent mean *d'* values, violins' width represents the frequency of *d'* values, and boxplots show distributions in quartiles. Discrimination abilities decrease linearly with less exposure and longer delays, and are impaired across the board for dissimilar pictures compared to similar pictures.

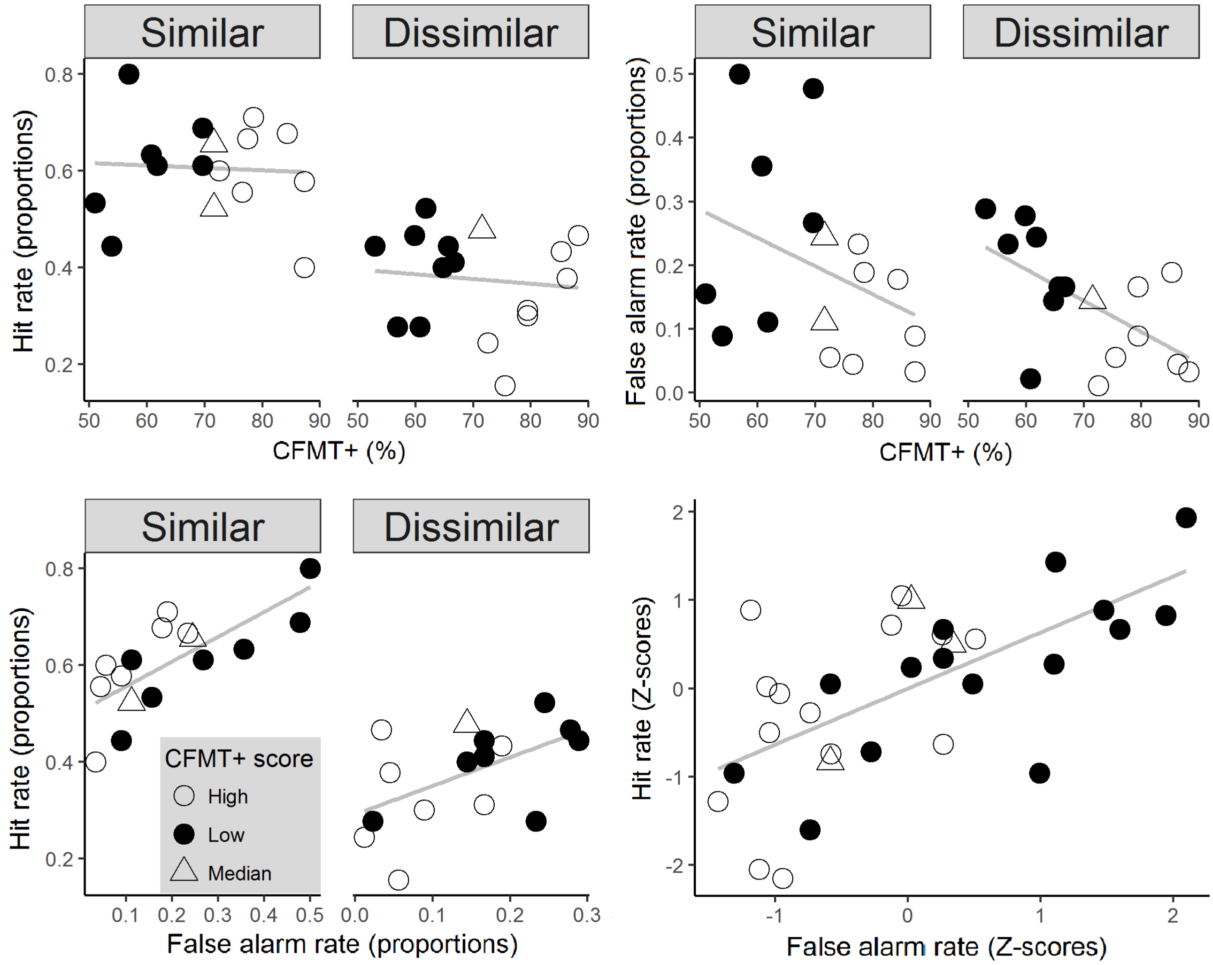


Figure 4. Top panels show correct recognition (hit) rates (left) and false alarm rates (right) in each picture similarity condition and as a function of individual CFMT+ scores (in percentage). Bottom panels show false recognition rates plotted against hit rates for each participant (per similarity condition on the left, and across participants on the right, in standardised Z-scores to compensate for differences in performance between the two conditions). Full circles show participants with low CFMT+ scores and open circles are those with high scores following a median split. Triangles show participants whose score is equal to the median. This set of figures shows that individual face recognition skills, as measured by a benchmark test, negatively correlate with false alarm rates, but do not correlate with hit rates. In other words, poor and good recognisers have similar ranges of hits, but poor recognisers commit more false alarms than good recognisers.

Accuracy of familiarity judgment in the GoT test. Associations between accuracy of familiarity judgments and different measures within our own test follow a very similar pattern, see **Table**

2 and **Figure S1**. Again, accuracy of familiarity judgments is driven by false alarm rates, *r* = -.557, *p* = .001, rather than by hit rates, *r* = .255, *p* = .16. Although the association between accuracy and hit rate might be underpowered, a Steiger's test again confirms that accuracy is significantly better predicted by false alarms rates than by hit rates, *Z* = -5.627, *p* < .001.

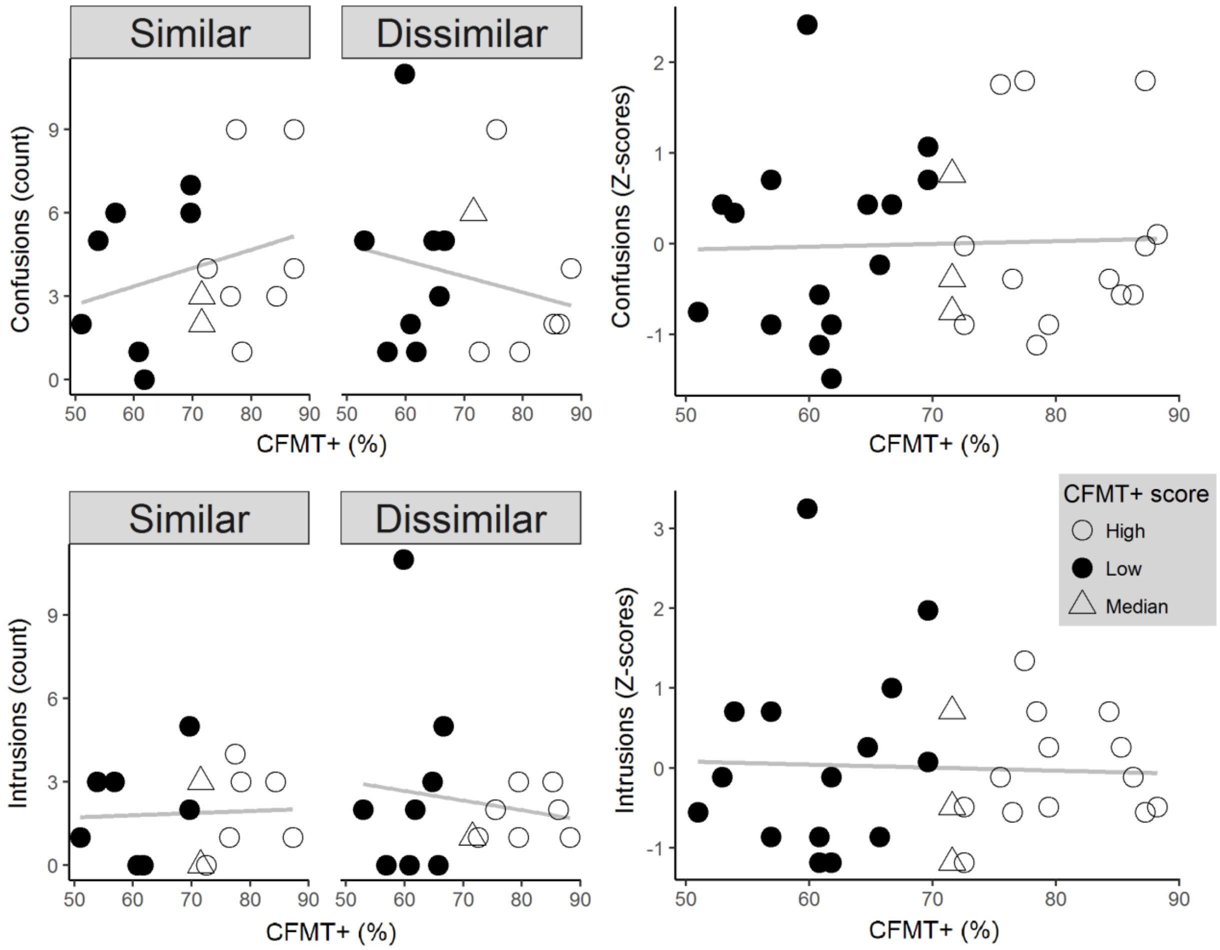


Figure 5. Misidentification errors as a function of participants' CFMT+ scores (in percentage). Top panels show confusions (i.e., identifications that do not match actors' identity) and bottom panels show intrusions (i.e., erroneous identification of strangers). Left panels show individual participants in each similarity condition (performance expressed in number of occurrences) and right panels show results across participants (in Z-scores). Analyses conducted across similarity conditions show that individual face recognition skills do not predict misidentification errors.

Better accuracy predicts a lower likelihood to report mere familiarity for strangers, *r* = -.555, *p* = .001, but not significantly so for actors, *r* = -.233, *p* = .199. More accurate recognisers are more likely to correctly identify, *r* = .693, *p* < .001, and to name the actors, *r* = .574, *p* = .001, than less accurate ones. However, again, how accurate people are overall does not significantly predict how likely they are to commit identification errors on actors, *r* = -.143, *p* = .434, and on strangers, *r* = -.219, *p* = .228.

Of note, although there are no significant associations between *c* (i.e., decision criterion) and accuracy, *r* = .259, *p* = .153, nor between *c* and CFMT+ scores, *r* = .335, *p* = .061, the bottom right panel of **Figure 4** and **Figure S1** suggest that superior recognisers might in general be more cautious in their decisions than poorer recognisers who either report few recognitions (hits *and* false alarms) or many, either correct or incorrect.

Confidence ratings. How well can people judge their own recognition performance? On average, people are more confident when they make correct familiarity judgments than when they are incorrect, *Z* = -4.787, *p* < .001, *r* = .846, see **Table 2**. At the individual level, there is no significant relationship between confidence ratings and accuracy on our *GoT* task or on the CFMT+, all *p*s > .1, except that people with higher CFMT+ scores tend to be more confident when they are correct, *r* = .339 , *p* = .06.

Discussion

We have described a highly challenging and ecologically-valid recognition task that controls exposure to a person’s face and delay since they were last seen. As predicted by many models of face recognition, performance increased with greater exposure (e.g., Kramer, Young, & Burton, 2018), and we also show that it decreased with longer delays. Despite these broadly consistent findings, closer analysis bring several important and novel insights to light. First, we find that recognising and identifying familiar faces is a far from trivial task. People experience surprising difficulties in recognising faces to which they have been repeatedly exposed, and struggle even more in retrieving their identity and their names. Importantly, even people with the best recognition skills forget faces or misidentify them. Second, we show that extra-facial features (like hair colour, beards, or accessories) make important contributions to recognition for both good and poor recognisers, even for the most familiar faces. Third, we show that decision processes may be more important for face recognition than has previously been appreciated. Specifically, we find that good performance is driven primarily by the rejection of strangers' faces, and not the detection of familiar ones. We expand on each of these insights below.

Exposure and delay. Although face recognition improved with greater exposure, performance was not as good as we might expect. Only the most prominent actors were correctly identified and named more often than they were just recognised; and identification errors (confusions and intrusions) occasionally occurred at all exposure levels. These findings are consistent with models of face recognition in which facial and semantic information are processed in different units (Bruce & Young, 1986; Burton et al., 1999; Gobbini & Haxby, 2007). Further, the least exposed actors were *never* correctly identified, and only rarely were they recognised, even if their brief appearance was sometimes shocking (e.g., one of them raped another character). Rates of feelings of mere familiarity were barely higher for these actors than for strangers, suggesting that reports of mere familiarity are poor indicators of a genuine encounter with a person; especially in some poorer recognisers who indiscriminately perceive many actors and strangers as familiar.

Recognition was similarly affected by delay, showing that humans forget even well-learned faces over time. Despite the presence of people with excellent face recognition skills in our sample, recognition performance is impaired by the mere passage of time, revealing human limitations in terms of degradation over time compared to automatic systems. This finding indicates that the claim made by some super-recognisers that they do not forget faces might be unfounded, and **Figure 4** shows that the highest hit rate was around 80% (by a person with a low CFMT+ score who also committed around 50% of false alarms).

Although screen time provided an objective measures of exposure, we recognise that it might be a proxy for other variables that co-vary with it, and which might constitute the underlying mechanisms that drive familiarisation. For example, frequently-encountered characters have stronger semantic representations based on their actions and associations. Associating a face with semantic information (Schwartz & Yovel, 2016) or specific abstract labels (e.g., names or even one letter, McGugin, Tanaka, Lebrecht, Tarr, & Gauthier, 2011) facilitates subsequent recognition. Less available semantic information for minor characters, who were probably named less often too, could thus contribute to poorer recognition performance alongside duration of exposure itself. This co-accumulation of visual exposure and semantic associations happens in the real world too, where we pick up facts about people over encounters, even if they are strangers (e.g., the places we see them, the type of clothes they wear, the type of job they might do). Another factor that likely co-varies with screen time is within-person variation, which is known to facilitate the learning of new faces (Baker et al., 2017; Menon, White, & Kemp, 2015; Ritchie & Burton, 2017). Presumably, more prominent actors were encountered under more variable conditions than less prominent ones: they were seen over a range of lighting conditions, in different contexts, under different camera angles, with various facial expressions, etc. Finally, across multiple encounters, characters might have gained different motivational or emotional values that could modulate recognition of their faces. These covariations are difficult to prevent in naturalistic learning conditions, and even in lab-based learning studies where, for example faces might become associated with episodic information or abstract labels (e.g., the man that looks like your neighbour) over time or across repetitions. The exact mechanisms whereby exposure leads to better recognition remain to be elucidated, and will likely require controlled laboratory studies to isolate specific causal factors.

Similarity and robustness of representations. Recognition of even familiar faces was surprisingly vulnerable to superficial changes in appearance: it was impaired by the type of changes that occur routinely in the real world, like a new beard, hair colour, or glasses. This result contradicts current models of familiarisation which posit that exemplars of a person's face are averaged over encounters, yielding a robust representation based on invariant facial features that allow recognition on subsequent encounters (Burton et al., 2005; Burton et al., 1999; Jenkins & Burton, 2011). Instead, it seems that even the most prominent actors' faces are not reliably recognised based on invariant facial features.

A possible explanation for this unexpected finding lies in the fact that *GoT* actors' faces have been learned by our participants in just one role, thereby displaying less within-person variation than some other celebrities might display, and therefore leading to less robust representations (Burton et al., 2016; Menon, White, & Kemp, 2015). However, the averaging process should have occurred for actors in *GoT* too, who have aged and altered their appearance across the six seasons. Moreover, more prominent actors must have been encountered under more variable conditions than less prominent ones and so recognition should have been impaired less by changes in appearance in main heroes and lead characters.

Alternatively, previous studies of face recognition have overestimated the robustness of familiar faces' representations and underestimated people's vulnerability to changes in appearance. Perhaps the importance of extra-facial information has been overlooked because in many studies external features are simply removed (see also Sinha & Poggio, 1996). Moreover, where extra-facial information is conserved, researchers might have tended, even unintentionally, to select only the most representative pictures of celebrities or personally familiar people in their tests, (i.e., pictures in which the person's appearance does not depart much from their usual appearance), thereby boosting recognition performance. In support of this idea, pictures of celebrities rated as displaying a good likeness are recognised more efficiently than pictures rated as displaying a poor likeness (Ritchie, Kramer, & Burton, 2018), and iconic images of celebrities like Marylin Monroe or Ernesto Che Guevara are easier to recognise than either altered versions of the same images or less-commonly seen pictures of the same people (Carbon, 2008). By deliberately seeking out photos with different extra-facial characteristics than the character, we were able to challenge recognition abilities and show how important such features are.

Interestingly, both good and poor recognisers seem to be impaired by these superficial changes, which led to fewer hits across the board (see top left panel of **Figure 4**). Future research should provide more robust support for this pattern with within-subject manipulations of similarity, making it possible to examine the impact of dissimilarity at the individual level. Anecdotally, misidentifications committed by both good and poor recognisers were often based on an over-reliance on superficial features like hair or the shape of the head (e.g., a bald foil was identified multiple times as a character who is bald in the show). Such over-reliance on extra-facial features is one of the hallmark symptoms of developmental prosopagnosia (Murray, Hills, Bennetts, & Bate, 2018). In fact, people with propopagnosia can perform in the normal range when hair and eyebrows allow them to distinguish individuals within a small set of novel faces (Duchaine & Weidenfeld, 2003). It could be that normal recognisers also succumb to such over-reliance in some contexts. The Clinton-Gore composite illusion (in which Clinton's internal features are unnoticeably pasted into Al Gore's face; Sinha & Poggio, 1996) might illustrate this bias toward extra-facial features in recognition. This idea is also consistent with reports by typical participants that they have sometimes failed to recognise even their own face in daily life, often because of unusual features like hairstyle (Brédart & Young, 2004).

Decisions processes. Changes in appearance not only impair recognition per se, but also affect decision criteria: in contexts where recognition is harder (i.e., the dissimilar condition), people are less likely to report recognition. This is true even for photos of actors who played bit parts and photos of strangers, which did not actually differ between similar and dissimilar conditions. By contrast, similarity did not affect reports of confidence. Although people were more confident when right than wrong, average levels of confidence still seem unrealistically high when judgments are in error (i.e., between 3.23 and 3.47 on a 5-point Likert scale, which translates into being somewhat confident). At the individual level, we did not find a significant relationship between confidence ratings and accuracy on our *GoT* task, and only a moderate relationship between confidence on correct trials and the CFMT+, suggesting that some people are overly confident regardless of their actual performance (see also Palermo et al., 2017).

These findings highlight the importance of decision processes in face recognition skills. We show that individual differences, measured via both recognition performance on our task *and* via a completely different test (i.e., CFMT+), are driven by vulnerability to false recognitions, significantly more than by accurate recognitions. Superior recognisers do not necessarily recognise more actors than others. However, they know better when they encounter a face for the first time, and are better able to reject strangers as unfamiliar. Likewise, when confronted with the most challenging trials of the CFMT+, those identified as superior recognisers might be better at correctly rejecting the two foils, and then proceed by elimination to pick the target[ii]. These conclusions are in line with applied research showing that good eyewitnesses are people who are able to reject a line-up when the perpetrator is absent (Bindemann, Brown, Koyas, & Russ, 2012). A closer look at recent data on super-recognisers' unfamiliar face matching abilities also suggests that they differ from controls most in false alarm rate (Bobak, Hancock, et al., 2016). Our results also show that superior recognisers are better able to place and name the people they recognise. Since they tend to be more conservative in their familiarity judgments than poorer recognisers, they might be more inclined to report recognition only when they can also remember the person's identity. Importantly, they still occasionally make identification errors just like anyone else, and these errors are not predicted by CFMT+ scores nor accuracy on our *GoT* task.

At the other end of the spectrum, poor recognisers struggle to place faces and tend to report mere feelings of familiarity following both correct and incorrect recognitions. This suggests that they have difficulty with decision processes when confronted with indiscriminate and pervasive feelings of familiarity. It could be that they fail to get the "big picture" and are

[ii] One participant with one of the highest CFMT+ scores spontaneously reported using that strategy.

misled by features (e.g., crooked nose, bald head) that are similar across different individuals (because we selected foils whose appearance broadly matched that of people in *GoT*), making them undiagnostic when used in isolation. This hypothesis requires further assessment, but is consistent with recent accounts of stronger holistic processing in good recognisers compared to poor ones (DeGutis, Wilmer, Mercado, & Cohan, 2013; Richler, Cheung, & Gauthier, 2011; Wang, Li, Fang, Tian, & Liu, 2012), delayed access to global relative to local shape information in developmental prosopagnosia (Gerlach, Klargaard, Petersen, & Starrfelt, 2017), reports of impaired holistic processing in acquired prosopagnosia (Busigny, Joubert, Felician, Ceccaldi, & Rossion, 2010; Ramon, Busigny, Gosselin, & Rossion, 2016), and the finding that, at the group level, successful familiarisation with novel faces is associated with less focus on local features (Ramon & Van Belle, 2016).

These observations have a parallel in research on false memories, which can be caused by source monitoring errors that arise when perceptually similar elements of true and false events lead to the erroneous conclusion that retrieving a specific element marks a memory as genuine (see Johnson, Raye, Mitchell, & Ankudowich, 2012, for a review). It is possible that poor face recognition abilities stem in part from similar meta-cognitive errors following feeling of familiarity elicited by isolated facial features that several people might share (see also the discussion in Richler, Floyd, & Gauthier, 2015 on how the repeated presentation of the same face parts in holistic processing tasks particularly affects associations with CFMT performance). A recent investigation of the misinformation effect (i.e., the development of false memories after exposure to misleading information) has revealed a weak but significant negative association between CFMT scores and the susceptibility to develop false memories (Zhu et al., 2010), and so it could be that poor performance exhibited with faces expands to other memory tasks. Finally, our findings here are also in line with recent reports that poorer recognisers show great variability in recognition strategy (see Bate & Tree, 2017; Esins, Schultz, Stemper, Kennerknecht, & Bülthoff, 2016; Palermo et al., 2017): here we see that while some are conservative and report few recognitions in general (i.e., few false alarms but also few hits), others indiscriminately report many recognitions for both actors and strangers. Better understanding of individual differences will require systematic examination of the many cognitive processes involved in person identification: perceptual discrimination, visual memory, semantic memory, naming, decision-making, and source monitoring.

Conclusions

Although our research was motivated by our desire to address theoretical questions regarding the process of familiarisation, our findings have important practical implications as well. While criminals know too well how simple disguises can help them escape prosecution, many innocents are convicted based on testimony by eyewitnesses who saw a perpetrator just once. Our findings mirror those in the field showing that recognitions based on brief encounters are likely to be wrong, and furthermore that confidence of a witness in these conditions does not predict accuracy (Morgan et al., 2007; Wells & Olson, 2003). We also show that superior recognisers might make useful contributions to law enforcement agencies, but they are not infallible, so their judgments too must be supported by corroborating evidence.

The novel approach developed here combines ecological validity and tight experimental control to capture the transition from stranger to familiar face. Findings from the rich dataset confirm that familiarisation is a slow and incremental process, but also show that humans are not as expert as we might think. Recognition of even highly familiar faces (the main heroes and lead characters here) is impaired by superficial changes in appearance, fades with time, and is dependent as much on decision processes as on face processing skills. Beyond its empirical contributions, our research also highlights the opportunities provided by virtual worlds like the *GoT* universe to study developmental processes in controlled but contextually rich environments.

Context of research

The idea for this research emerged from CD's will to find a way to really challenge face recognition skills of people with superior abilities. She has been researching face processing for over a decade, studying attentional biases and the impact of artistic expertise, emotions or personal relevance. GG's work on face processing has primarily focused on emotional expression, but this collaboration has inspired an interest in person recognition. We are currently developing a program of research that will assess new hypotheses outlined here. We are particularly interested in moving face recognition research away from a focus on perceptual processing of facial features in isolation, and toward a richer perspective that incorporates extra-facial features and captures the full range of cognitive processes that might account for the wide range of individual differences in ability.

Authors' notes

Contributions: CD developed the study concept and design, prepared material, collected and analysed data, and drafted the manuscript. AW helped with material preparation, data collection, and data encoding, and approved the final version of the manuscript. GG contributed to the study design, data analyses, and manuscript writing.

The exhaustive list of actors/characters presented in the *GoT* test are available as Supplementary Material. Summary data and an overview of stimulus material are visible at [https://osf.io/xjmzp/ and https://osf.io/wg8vx/]. We thank Catherine Barsics for her comments on the methods, Rebecca Hamilton for her help with pilot data collection, and David White, Richard Kemp, Anne MacAskill, and Raymond Klein for their comments on the data. Portions of the data were presented at the Experimental Psychology Conference (2017) and at the European Conference on Visual Perception (2017). A preliminary version of this manuscript was posted on the Open Science Framework website in November 2017 [https://osf.io/nc6s8/].

References

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty Years of Memory for Names and Faces: A Cross-Sectional Approach. *Journal of Experimental Psychology: General*, *104*(1), 54–75. http://doi.org/10.1037/0096-3445.104.1.54

Baker, K. A., Laurence, S., & Mondloch, C. J. (2017). How does a newly encountered face become familiar? The effect of within-person variability on adults' and children's perception of identity. *Cognition*, *161*, 19–30. http://doi.org/10.1016/j.cognition.2016.12.012

Behrmann, M., & Avidan, G. (2005). Congenital prosopagnosia: face-blind from birth. *Trends in Cognitive Sciences*, *9*(4), 180–187. http://doi.org/10.1016/j.tics.2005.02.011

Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012). Individual differences in face identification postdict eyewitness accuracy. *Journal of Applied Research in Memory and Cognition*, *1*(2), 96–103. http://doi.org/10.1016/j.jarmac.2012.02.001

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

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting Superior Face Recognition Skills in a Large Sample of Young British Adults. *Frontiers in Psychology*, *7:1378*. http://doi.org/10.3389/fpsyg.2016.01378

Brédart, S., & Devue, C. (2006). The accuracy of memory for faces of personally known individuals. *Perception*, *35*(1), 101–106. http://doi.org/10.1068/p5382

Brédart, S., & Young, A. W. (2004). Self-recognition in everyday life. *Cognitive Neuropsychiatry*, *9*(3), 183–97. http://doi.org/10.1080/13546800344000075

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*(4), 339–360. http://doi.org/10.1037/1076-898X.5.4.339

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305–327. http://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Burke, D., Taubert, J., & Higman, T. (2007). Are face representations viewpoint dependent? A stereo advantage for generalising across different views of faces. *Vision Research*, *47*(16), 2164–2169. http://doi.org/10.1016/j.visres.2007.04.018

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1467–1485. http://doi.org/10.1080/17470218.2013.800125

Burton, A. M., Bruce, V., & Hancock, P. J. B. (1999). From pixels to people: A model of familiar face recognition. *Cognitive Science*, *23*(1), 1–31. http://doi.org/10.1016/S0364-0213(99)80050-0

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: the power of averages. *Cognitive Psychology*, *51*(3), 256–84. http://doi.org/10.1016/j.cogpsych.2005.06.003

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, *40*(1), 202–223. http://doi.org/10.1111/cogs.12231

Busigny, T., Joubert, S., Felician, O., Ceccaldi, M., & Rossion, B. (2010). Holistic perception of the individual face is specific and necessary: evidence from an extensive case study of acquired prosopagnosia. *Neuropsychologia*, *48*(14), 4057–92. http://doi.org/10.1016/j.neuropsychologia.2010.09.017

Carbon, C. C. (2008). Famous faces as icons. The illusion of being an expert in the recognition of famous faces. *Perception*, *37*(5), 801–806. http://doi.org/10.1068/p5789

Carey, S., & Diamond, R. (1977). From piecemeal to configural represention of faces. *Science*, *195*(4275).

Clutterbuck, R., & Johnston, R. a. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception*, *31*(8), 985–994. http://doi.org/10.1068/p3335

DeGutis, J., Wilmer, J., Mercado, R. J., & Cohan, S. (2013). Using regression to measure holistic face processing reveals a strong link with face recognition ability. *Cognition*, *126*(1), 87–100. http://doi.org/10.1016/j.cognition.2012.09.004

Devue, C., Collette, F., Balteau, E., Degueldre, C., Luxen, A., Maquet, P., & Brédart, S. (2007). Here I am: the cortical correlates of visual self-recognition. *Brain Research*, *1143*(1), 169–82. http://doi.org/10.1016/j.brainres.2007.01.055

Duchaine, B. C., & Weidenfeld, A. (2003). An evaluation of two commonly used tests of unfamiliar face recognition. *Neuropsychologia*, *41*(6), 713–720. http://doi.org/10.1016/S0028-3932(02)00222-1

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external feature : Some implications for theories of face recognition. *Perception*, *8*, 431–439. http://doi.org/10.1068/p080431

Ewbank, M. P., & Andrews, T. J. (2008). Differential sensitivity for viewpoint between familiar and unfamiliar faces in human visual cortex. *NeuroImage*, *40*(4), 1857–1870. http://doi.org/10.1016/j.neuroimage.2008.01.049

Freire, A., & Lee, K. (2001). Face Recognition in 4- to 7-Year-Olds: Processing of Configural, Featural, and Paraphernalia Information. *Journal of Experimental Child Psychology*, *80*(4), 347–371. http://doi.org/10.1006/jecp.2001.2639

Freiwald, W., Yovel, G., & Duchaine, B. (2016). Face Processing Systems: From Neurons to Real World Social Perception. *Annual Review of Neuroscience*, *39*, 325–346. http://doi.org/10.1146/annurev-neuro-070815-013934

Ge, L., Luo, J., Nishimura, M., & Lee, K. (2003). The lasting impression of Chairman Mao: hyperfidelity of familiar-face memory. *Perception*, *32*(5), 601–614. http://doi.org/10.1068/p5022

Gerlach, C., Klargaard, S. K., Petersen, A., & Starrfelt, R. (2017). Delayed processing of global shape information in developmental prosopagnosia. *PLoS ONE*, *12*(12), 1–20. http://doi.org/10.1371/journal.pone.0189253

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia*, *45*(1), 32–41. http://doi.org/10.1016/j.neuropsychologia.2006.04.015

Hill, H., Schyns, P. G., & Akamatsu, S. (1997). Information and viewpoint dependence in face recognition. *Cognition*, *62*(2), 201–22. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9141907

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

Johnston, A., Hill, H., & Carman, N. (1992). Recognising faces: Effects of lighting direction, inversion, and brightness reversal. *Perception*, *21*(3), 365–375. http://doi.org/10.1068/p210365n

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: a review. *Memory*, *17*(5), 577–596. http://doi.org/10.1080/09658210902976969

Kramer, R. S. S., & Ritchie, K. A. Y. L. (2016). Disguising Superman: How glasses affect unfamiliar face matching. *Applied Cognitive Psychology*, *30*(6), 841–845. http://doi.org/10.1002/acp.3261

Kramer, R. S. S., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. *Cognition*, *172*(June 2017), 46–58. http://doi.org/10.1016/j.cognition.2017.12.005

McGugin, R. W., Tanaka, J. W., Lebrecht, S., Tarr, M. J., & Gauthier, I. (2011). Race-specific perceptual discrimination improvement following short individuation training with faces.

Cognitive Science, *35*(2), 330–347. http://doi.org/10.1111/j.1551-6709.2010.01148.x

Menon, N., White, D., & Kemp, R. I. (2015). Variation in photos of the same face drives improvements in identity verification. *Perception*, *44*(11), 1332–1341. http://doi.org/10.1177/0301006615599902

Morgan, C. A., Hazlett, G., Baranoski, M., Doran, A., Southwick, S., & Loftus, E. (2007). Accuracy of eyewitness identification is significantly associated with performance on a standardized test of face recognition. *International Journal of Law and Psychiatry*, *30*(3), 213–223. http://doi.org/10.1016/j.ijlp.2007.03.005

Murray, E., Hills, P. J., Bennetts, R. J., & Bate, S. (2018). Identifying hallmark symptoms of developmental prosopagnosia for non-Experts. *Scientific Reports*, *8*(1), 1–12. http://doi.org/10.1038/s41598-018-20089-7

Natu, V., O'Toole, A. J., & Toole, A. J. O. (2011). The neural processing of familiar and unfamiliar faces : A review and synopsis. *British Journal of Psychology*, *102*(4), 726–747. http://doi.org/10.1111/j.2044-8295.2011.02053.x

Palermo, R., Rossion, B., Rhodes, G., Laguesse, R., Hall, B., Albonico, A., … Mckone, E. (2017). Do people have insight into their face recognition abilities ? *The Quarterly Journal of Experimental Psychology*, *70*(2), 218–233. http://doi.org/10.1080/17470218.2016.1161058

Pilz, K. S., Thornton, I. M., & Bulthoff, H. H. (2006). A search advantage for faces learned in motion. *Experimental Brain Research*, *171*(4), 436–447. http://doi.org/10.1007/s00221-005-0283-8

Ramon, M., Busigny, T., Gosselin, F., & Rossion, B. (2016). All new kids on the block ? Impaired holistic processing of personally familiar faces in a kindergarten teacher ... *Visual Cognition*, *24*(5-6), 321–355. http://doi.org/10.1080/13506285.2016.1273985

Ramon, M., & Van Belle, G. (2016). Real-life experience with personally familiar faces enhances discrimination based on global information. *PeerJ*, *4*, e1465. http://doi.org/10.7717/peerj.1465

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011). Holistic processing predicts face recognition. *Psychological Science*, *22*(4), 464–71. http://doi.org/10.1177/0956797611401753

Richler, J. J., Floyd, R. J., & Gauthier, I. (2015). About-face on face recognition ability and holistic processing. *Journal of Vision*, *15*(9), 15. http://doi.org/10.1167/15.9.15

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *The Quarterly Journal of Experimental Psychology*, *70*(5), 1–9. http://doi.org/10.1080/17470218.2015.1136656

Ritchie, K. L., Kramer, R. S. S., & Burton, A. M. (2018). What makes a face photo a “good likeness”? *Cognition*, *170*(April 2017), 1–8. http://doi.org/10.1016/j.cognition.2017.09.001

Ritchie, K. L., Smith, F. G., Jenkins, R., Bindemann, M., White, D., & Burton, A. M. (2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. *Cognition*, *141*, 161–169. http://doi.org/10.1016/j.cognition.2015.05.002

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: people with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252–7. http://doi.org/10.3758/PBR.16.2.252

Schwartz, L., & Yovel, G. (2016). The roles of perceptual and conceptual information in face recognition. *Journal of Experimental Psychology: General*, *145*(11), 1493–1511. http://doi.org/10.1037/xge0000220

Sinha, P., & Poggio, T. (1996). I think I know that face... *Nature*, *384*(6608), 404–404. http://doi.org/10.1038/384404a0

Susilo, T., & Duchaine, B. (2013). Advances in developmental prosopagnosia research. *Current Opinion in Neurobiology*, *23*(3), 423–429. http://doi.org/10.1016/j.conb.2012.12.011

Toseeb, U., Keeble, D. R. T., & Bryant, E. J. (2012). The significance of hair for face recognition. *PLoS ONE*, *7*(3), 1–8. http://doi.org/10.1371/journal.pone.0034144

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual differences in holistic processing predict face recognition ability. *Psychological Science*, *23*(2), 169–177. http://doi.org/10.1177/0956797611420575

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology*, *54*, 277–295. http://doi.org/10.1146/annurev.psych.54.101601.145028

Young, A. W., & Burton, A. M. (2017). Recognizing Faces. *Current Directions in Psychological Science*, *26*(3), 212–217. http://doi.org/10.1177/0963721416688114

Zhu, B., Chen, C., Loftus, E. F., Lin, C., He, Q., Chen, C., … Dong, Q. (2010). Individual differences in false memory from misinformation: Cognitive factors. *Memory*, *18*(5), 543–555. http://doi.org/10.1080/09658211.2010.487051

Supplementary Material

List of persons selected in each delay and exposure conditions (actor name / character name)

Season 6 – Main heroes

Alfie Allen / Theon Greyjoy; Nikolaj Coster-Waldau / Jaime Lannister; Kit Harington / Jon Snow; Lena Headey / Cersei Lannister; Sophie Turner / Sansa Stark; Maisie Williams / Arya Stark

Season 6 – Lead characters

Ben Crompton / Eddison Tollett; Liam Cunningham / Davos Seaworth; Jerome Flynn / Bronn; Conleth Hill / Lord Varys; Finn Jones / Loras Tyrell; Gemma Whelan / Yara Greyjoy

Season 5 – Lead characters

Ian Beattie / Meryn Trant; Charles Dance / Tywin Lannister; Stephen Dillane / Stannis Baratheon; Ciarán Hinds / Mance Rayder; Ian McElhinney / Barristan Selmy

Season 4 – Lead characters

Jack Gleeson / Joffrey Baratheon; Sibel Kekilli / Shae; Rose Leslie / Ygritte; Pedro Pascal / Oberyn Martell; Mark Stanley / Grenn

Season 3 – Lead characters

Esmé Bianco / Ros; Oona Chaplin / Talisa Maegyr; James Cosmo / Jeor Mormont; Joe Dempsie / Gendry; Michelle Fairley / Catelyn Stark; Richard Madden / Robb Stark

Season 1-2 – Lead characters

Mark Addy / Robert Baratheon; Gethin Anthony / Renly Baratheon; Harry Lloyd / Viserys Targaryen; Jason Momoa / Khal Drogo; Donald Sumpter / Maester Luwin

Season 6 – Support characters

Roger Ashton-Griffiths / Mace Tyrell; Charlotte Hope / Myranda; Anton Lesser / Qyburn; Faye Marsay / Waif; Eugene Simon / Lancel Lannister; Indira Varma / Ellaria Sand

Season 5 – Support characters

Tara Fitzgerald / Selyse Baratheon; Joel Fry / Hizdahr zo Loraq; Kerry Ingram / Shireen Baratheon; Will Tudor / Olyvar; Peter Vaughan / Maester Aemon

Season 4 – Support characters

Josef Altin / Pypar; Kate Dickie / Lysa Arryn; Burn Gorman / Karl Tanner; Noah Taylor / Locke; Tony Way / Dontos Hollard

Season 3 – Support characters

Mackenzie Crook / Orell; Philip McGinley / Anguy; Jamie Michie / Steelshanks Walton; Robert Pugh / Craster; John Stahl / Rickard Karstark

Season 1-2 – Support characters

Amrita Acharia / Irri; Ron Donachie / Rodrik Cassel; Ralph Ineson / Dagmer Cleftjaw; Francis Magee / Yoren; Roxanne McKee / Doreah; Dar Salim / Qotho

Season 6 – Bit parts

Matteo Elezi / Young Benjen; Angelique Fernandez / Dothraki Widow; Kevin Horsham / Westerosi Trader; Eddie Jackson / Belicho Paenymion; Fergus Leathem / Young Rodrik; Brenden O'Rourke / Mummer 'Background Panel

Season 5 – Bit parts

Ian Lloyd Anderson / Derek; Morgan C. Jones / Braavosi Captain; Stella McCusker / Old Woman; James McKenzie Robinson / Joss; Lacy Moore / Braavosi Madam; Rebecca Scott / The Maiden

Season 4 – Bit parts

Jazzy De Lisser / Tansy; Jody Halse / Adrack Humble; Karl Jackson / Unsullied; Rhodri Miles / First Mate; Deirdre Monaghan / Morag; Lois Winstone / Molestown Whore

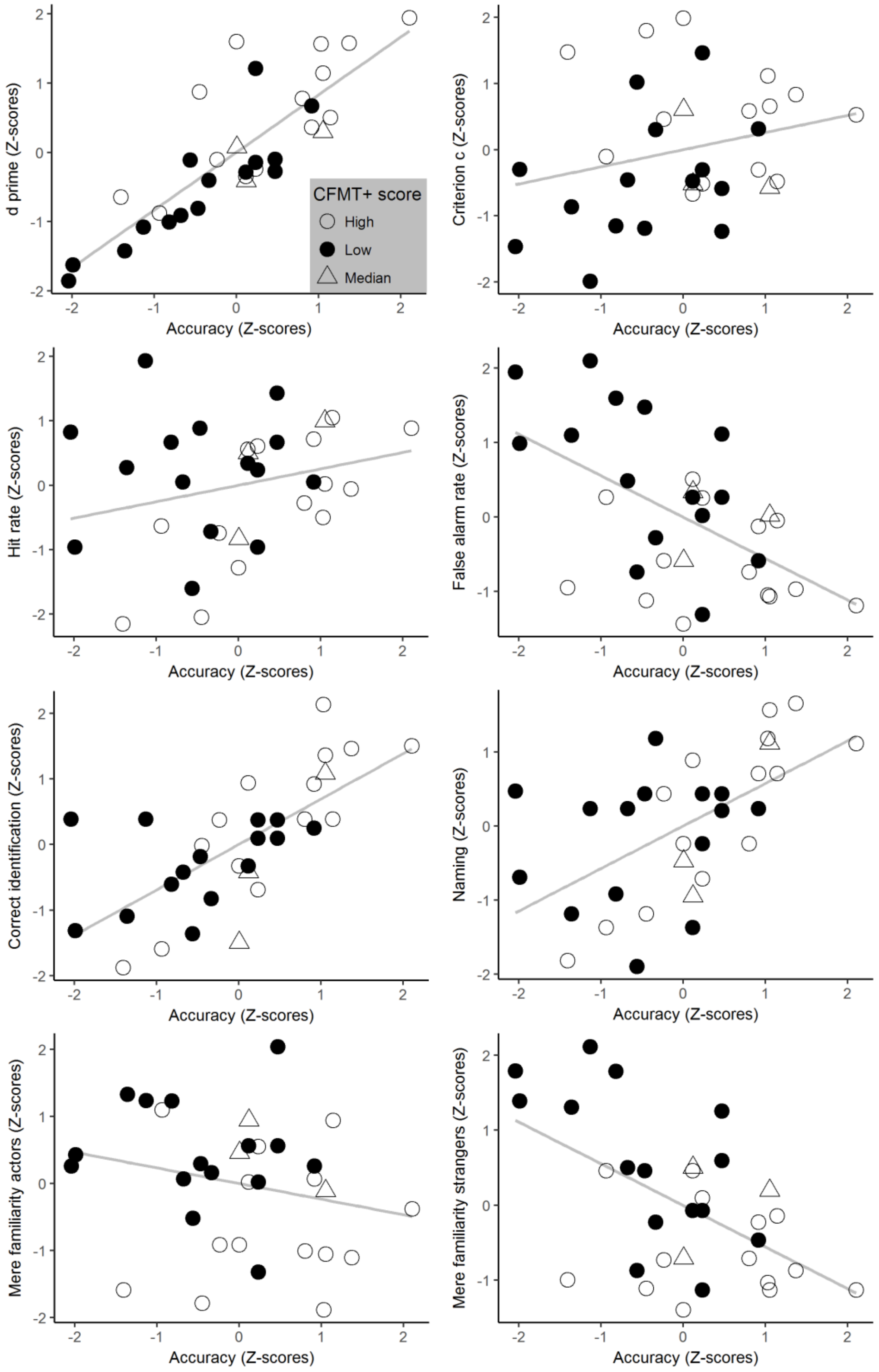
Season 3 – Bit parts

Cliff Barry / Greizhen mo Ullhor; Shaun Blaney / Karstark Lookout; Oddie Braddell / Wendel Manderly; Grace Hendy / Merry Frey; Aisling Jarrett-Gavin / Margaery's Handmaid; Michael Shelford / Torturer

Season 1-2 – Bit parts

Peter Ballance / Farlen; David Coakley / Drennan; Mark Coney / Lord Galbart Glover; Rhodri Hosking / Mycah; Susie Kelly / Masha Heddie; Steve Wilson / Theon's Master of Hounds

Figure S1. Associations between accuracy on the *GoT* test and other measures (from top left to bottom right: d’, criterion c, hit rate, false alarm rate, total correct identification rate, naming rate, rates of mere familiarity for actors and rates of mere familiarity for strangers; all expressed in Z-scores). Individual participants are depicted as a function of their CFMT+ scores (low, median, or high).


d prime (Z-scores)
Criterion c (Z-scores)
Hit rate (Z-scores)
False alarm rate (Z-scores)
Correct identification (Z-scores)
Naming (Z-scores)
Mere familiarity actors (Z-scores)
Mere familiarity strangers (Z-scores)
Accuracy (Z-scores)
CFMT+ score
○ High
● Low
△ Median