Behavior Research Methods
2008, 40 (3), 840-857
doi: 10.3758/BRM.40.3.840

Toward a comprehensive test battery for face cognition: Assessment of the tasks

Grit Herzmann
Humboldt University Berlin, Berlin, Germany

Vanessa Danthiir
Humboldt University Berlin, Berlin, Germany
and the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Adelaide, Australia

and

Annekathrin Schacht, Werner Sommer, and Oliver Wilhelm
Humboldt University Berlin, Berlin, Germany

Despite the importance of face recognition in everyday life and frequent complaints about its failure, there is no comprehensive test battery for this ability. As a first step in constructing such a battery, we present 18 tasks aimed at measuring face perception, face learning, face recognition, and the recognition of facially expressed emotions. A sample of 153 healthy young adults completed all tasks. In general, reaction time measures showed high estimates of internal consistency; tasks focused on performance accuracy yielded reliabilities that were somewhat lower, yet high enough to support their use in a battery of face cognition measures. Some of the tasks allowed computation of established experimental effects in a within-subjects design, such as the part–whole effect. Most of these experimental effects were confirmed in our large sample, and valuable effect size estimates were obtained. However, in many cases these difference measures showed poor estimates of internal consistency.

Faces convey important socially relevant information, such as age, gender, emotional and other expressions, mate and social attractiveness, lip speech, and gaze direction. It is therefore crucial for successful interpersonal interaction to correctly perceive, learn, understand, and recognize the information that faces provide. There are enormous differences between people in face cognition, ranging from prosopagnosia, in which the learning and recognizing of new faces are highly impaired, to astonishing cases of memory for faces over many years. There is, however, no comprehensive multivariate battery to assess individual differences in face cognition. Such a test battery would lead to a better understanding of how faces are perceived, learned, and recognized, as well as providing insights into our social functioning in everyday life. The aim of the present article is to describe a collection of tasks to assess vital aspects of face cognition that are essential for a comprehensive test battery for face cognition. Here, we will first discuss prominent theoretical models and important conceptual distinctions in face cognition, before briefly reviewing the measures currently available and discussing the criteria necessary for tests meant to explore face cognition. In the main part of the article, 18 tasks and their psychometric characteristics will be described.

Theoretical Models and Conceptual Distinctions

The systems and processes involved in perceiving and recognizing familiar faces are captured by several models focusing on functions (Breen, Caine, & Coltheart, 2000; Bruce & Young, 1986; Burton, Bruce, & Hancock, 1999; Burton, Bruce, & Johnston, 1990; Ellis & Lewis, 2001) or describing underlying neuroanatomical substrates (Gobbini & Haxby, 2007; Haxby, Hoffman, & Gobbini, 2000). The various subprocesses captured by these models should be taken into account when developing an instrument for assessing individual differences in face cognition. These subprocesses will be briefly outlined below.

In order to recognize a person, one must be able, first, to identify facial structures as being, for example, eyes or a nose or mouth; next, to represent their spatial relations; and finally, to realize that these form a composite “face” object. In terms of face cognition models, *perceiving faces* implies structural encoding—that is, extracting pictorial and structural codes from faces and maintaining them for a short period of time. In general, when a face is seen, *pictorial codes* are derived during the perceptual processing of the retinal input. These codes are relatively raw images that contain much information irrelevant for face memory. Following the derivation of pictorial codes, viewpoint- and expression-independent descriptions (i.e.,

G. Herzmann, grit.herzmann@cms.hu-berlin.de

structural codes) of the viewed face are extracted. Structural codes have been suggested to mediate recognition of familiar faces, because they incorporate the facial features and their specific arrangement (configuration), a process required for distinguishing faces from each other. The important aspects of this high-level visual processing of faces are considered to be (1) the analysis of first-order features (i.e., facial elements that can be referred to in relative isolation, such as the size and shape of the nose or mouth), (2) the analysis of second-order or configurational features (i.e., spatial relationships among first-order features, such as the distance between the nose and mouth), and (3) the holistic perception and representation of faces (i.e., the perception and representation of faces as wholes or gestalts; Farah, Wilson, Drain, & Tanaka, 1998).

For each familiar face, the models postulate the existence of a *face recognition unit* (FRU), an interconnected set of structural codes stored within long-term memory. A viewed face is recognized as familiar when the structural codes derived during visual perception match those stored within the corresponding FRU. *Recognizing faces* thus requires the maintenance of structural codes stored in the FRUs, the comparison of stored and currently perceived facial structures, the correct reactivation of an FRU, and the familiarity decision process. Traditionally, face recognition is tested with already-familiar faces (such as celebrities, friends, etc.; see, e.g., Bruce & Valentine, 1985; Bruce & Young, 1986). Using such faces has important drawbacks, though, including differences in the frequency and duration of exposure to the faces, as well as in the availability and type of additional information, between both items and participants; all of these factors cause substantial construct-irrelevant variance in test performance. Ideally, initially unfamiliar faces that are learned within an experimentally controlled setting prior to testing should be used. The downside of using newly learned faces is the lack of many properties associated with faces that have been familiar over many years. Nevertheless, it is likely that newly learned faces not only allow superior control, as compared with the faces of celebrities, but also the capture of essential aspects of face cognition abilities.

All models of face cognition presuppose already-established facial representations and do not include the acquisition of such knowledge. It must, however, be assumed that variability in learning new faces will also constitute a source of individual differences in recognition of faces. Therefore, in our attempt to develop an instrument to measure individual differences in face cognition, we also considered the learning of new faces. As a face becomes familiar, external features (such as hairstyle) decline in importance for recognition, whereas internal features (such as mouth, nose, and eyes, as well as their relationships) become more salient (Bonner, Burton, Jenkins, McNeill, & Bruce, 2003). This finding suggests that the initial representations in memory for to-be-learned faces will later be replaced, at least to some extent, by greater emphasis on internal features. As yet, relatively little is known about the transition of a face from unfamiliar to familiar. Here, we circumvented the issue of the differential importance of internal and external features at different stages of face cognition by using only portraits in which external features were minimized.

For the course of face cognition following FRU activation, the relevant models suggest the activation of semantic (biographical) information and name codes. Because we focused on face perception and access to newly learned structural representations, we will not dwell on this issue here.

Apart from the face identification route, described above, all models postulate the separate processing of *changeable face codes*, such as the analysis of facially expressed emotions. Expression analysis comprises perceiving, analyzing, and identifying *facially expressed emotions*. By analyzing subtle changes in the internal features of a face and matching the perceived pattern to the stored pattern of facially expressed emotions, we can categorize a person as looking happy, sad, disgusted, and so forth.

Available Measures and Criteria for Indicators

Several tests are available for assessing individual differences in face cognition. Both the older test of Benton and Van Allen (1968) and Warrington's (1984) Recognition Memory Test have been criticized for also including face-irrelevant information, which could be used to accomplish the tasks by employing feature-based processing or general picture recognition skills (see, e.g., Dingle, Duchaine, & Nakayama, 2005; Duchaine & Nakayama, 2006). The more recent Bielefelder Famous Faces Test (BFFT; German adaptation by Fast, Fujiwara, & Markowitsch, 2005) uses already-familiar faces as stimuli. As discussed above, this approach is flawed, because the measure expresses substantial unwanted variance in test performance, due to differences in prior knowledge. Like the BFFT, the Cambridge Face Perception Task (CFPT; Dingle et al., 2005) and Face Memory Task (CFMT; Duchaine & Nakayama, 2006) are single tests that tap the wide field of face processing only in a very specific way. Their isolated application contains an unknown proportion of test-specific variance. In order to gain an adequate understanding of individuals' face-processing abilities, it is necessary to consider performance on a broader battery of indicators for these abilities. Furthermore, most tests focus solely on accuracy but neglect performance aspects related to the speed of face processing.

From a theoretical point of view, we hold the following four criteria as essential for constructing a comprehensive test battery for face cognition that can be used in scientific, clinical, or applied areas. First, each task must predominantly assess aspects of face cognition rather than picture or object processing. The intention is to assess abilities related to face cognition (e.g., holistic and configural processing of faces) but not to everyday visual recognition abilities (e.g., feature matching with symbols). We stress the difference between face and object recognition because neuropsychological studies of patients suffering from prosopagnosia (e.g., Duchaine & Nakayama, 2005; Farah et al., 1998), as well as experimental data from healthy participants (e.g., R. Diamond & Carey, 1986), provide compelling evidence that the processing of faces and nonface objects relies, at least partly, on different mechanisms (e.g., Farah et al., 1998). As a consequence, a comprehensive test battery should use face stimuli that show only internal features, and not hairstyle, adornments, clothing, or background, which could guide matching or recognition processes. In addition, when faces have to be matched, they should not be presented simultaneously for unlimited time, as in the Benton and Van Allen (1968) test, because this increases the likelihood and ease of using such strategies as feature matching.

Second, using already-familiar faces (e.g., celebrities) to test face recognition also brings variance that is irrelevant to face processing into the test performance. Thus, faces learned under experimental control should be used as the familiar stimuli.

Third, because individual differences may exist both in the accuracy and in the speed of face cognition, a comprehensive test battery should investigate performance in terms of both accuracy and speed, for each aspect of face cognition. Ideally, this distinction should be anchored in the type of task employed. The rationale for speeded tasks is that the trials in the tasks are so easy that every participant from the population for which the test is intended would be able to answer correctly on all trials if given enough time. The relevant source of individual variance in such tasks is therefore the speed of response. Conversely, in accuracy tasks, there are large differences in the quality of performance, even given unlimited time. Here, the source of individual variance is thus the accuracy of response.

Fourth, in a comprehensive battery, every aspect of face cognition should be assessed by more than a single test. Ideally, at least three tasks apiece within the accuracy- and speed-related approaches should probe each aspect of face cognition. These tasks should be compiled in a multivariate test battery. Obviously, in many practical applications, it will not be possible to use all tasks. However, the multivariate consideration of face-processing abilities, together with the available data, would allow for selecting optimal combinations of face-processing tests for specific practical measurement situations.

As a first step toward constructing a comprehensive test battery for face cognition in line with these criteria, we assembled a range of tasks aimed at probing the perception, learning, and recognition of faces, as well as the recognition of facially expressed emotions. The tasks were derived from or based on the experimental literature on face cognition and were chosen for their theoretical meaningfulness and the strength of their effects. We also included some published tests of specific face cognition abilities. In order to assess the suitability of the tasks and their psychometric properties, data were obtained from young adults. In addition, for some of the tasks we also report the experimental effects, their effect size (as expressed by Cohen's *d*; Cohen, 1988), and their psychometric properties, although these data are of secondary interest for the present purpose. These results are not only of interest in the context of test construction, but also provide a resource for experimental findings from an unusually large participant sample and according to the standards of experimental investigations. In the following sections, we first detail the general method of the study and then describe each of the 18 tasks, including its conceptual framework and procedure, as well as a summary of the most relevant results.

GENERAL METHOD

Participants

A total of 153 young adults (80 female), with an average age of 24.0 years ($SD = 4.5$), completed the tasks. Participants were recruited via newspaper ads, posters in various institutions, and radio broadcasts. They had to be between 18 and 35 years of age and Caucasian in order to meet the study's inclusion criteria. Being Caucasian was an inclusion criterion because we wanted to investigate the processing of Caucasian in comparison with Asian faces (see Tasks 4 and 9).[1]

Stimuli and Apparatus

Photographic portraits with neutral expressions were used as the practice and target stimuli; in the tasks assessing recognition of emotions, portraits displaying one of six emotions were also used. Apart from the stimuli in Tasks 4 and 9, all of the pictures displayed Caucasian faces. Unless specified otherwise, no face was used twice. Trials with female and male faces were balanced in number for all tasks, with the exception of the CFMT (Duchaine & Nakayama, 2006). If a trial involved the presentation of two faces—for example, in simultaneous matching—only faces of the same sex were used.

Photographs were obtained from several different resources (see the task descriptions for details). All portraits were converted to grayscale and were edited to the same format, with the aim of eliminating cues that could lead to correct task performance without reliance on face-specific skills. Thus, background and external facial features (clothing, hair, and ears) were excluded by fitting all portraits into vertical ellipses 200 pixels wide × 300 pixels high, corresponding to 5.1 × 7.6 cm. Only internal features of the faces up to the hairline were visible (see Figures 1–3 for examples). All faces were depicted in the frontal view, and also in three-quarter view for Tasks 3 and 5. Furthermore, only portraits that did not display distinct features or adornments (such as glasses, moles, beards, obvious makeup, or facial marks) were used.

The tasks were presented to the participants on 17-in. computer screens with refresh rates of 85 Hz and at a viewing distance of approximately 50 cm. All tasks, except the CFMT, were programmed using Inquisit 2.0 (2006); the CFMT is available in a Java-based version from Duchaine and Nakayama (2006).

Procedure

The study consisted of 18 tasks, assigned in the following way to the processes of interest: face perception (7 tasks), face learning (4 tasks), face recognition (4 tasks), and recognition of facially expressed emotions (3 tasks). Even though the tasks were designed to assess either speed or accuracy, participants were always instructed to respond as quickly and accurately as possible. Table 1 presents information on each task: the measurement intention (i.e., the aspect of face processing that was targeted), the performance parameter of prime interest (speed or accuracy), position in the sequence of this study, duration, and the numbers of trials and stimuli. The task sequence alternated between speed and accuracy tasks, and also between the different focal aspects of face cognition.

The tasks were administered by a trained proctor to groups of up to 9 participants, with each testing session lasting approximately 4 h, including breaks of 10 min after every hour of testing. No time limits were imposed for any task. Participants in each group worked in parallel on the same tasks, on individual yet similar computers. Except during Tasks 11 and 18, participants responded by pressing one of two buttons that were situated on the right and left sides of the computer keyboard. Participants were instructed to use their left and right index fingers and to always keep them above the relevant keys.

Each task began with an instruction page being presented on the screen, followed by about 10 practice trials, in which participants received trial-by-trial feedback about accuracy. The stimuli for practice trials were provided by Carbon and Leder (2007). After the practice trials, any participant queries were addressed, and then the experimental trials began. No feedback was provided for the experimental trials. Except for the CFMT, in all tasks an interstimulus interval of 1,300 msec was applied.

Unless specified otherwise in the individual task descriptions, the following procedure applied for every task. Each trial was presented without time limit and began with the presentation of a fixation cross for 200 msec, followed by the experimental stimuli, which remained on the screen until a response was made. All conditions and relevant aspects of the stimuli were balanced as evenly as possible across trials and other conditions (e.g., position of target face, gender of face, number of faces per condition, etc.). All tasks had a pseudorandomized trial sequence that was kept constant across participants.

Data Analysis

According to the measurement intention for each task, either speed or accuracy data were focused on as the main performance indicator. We examined the other performance indicator as well in order to more fully understand performance on the task. In some tasks, the difference between conditions was meant to reflect the cost/benefit of a specific face cognition function (e.g., the part–whole effect; Tanaka & Farah, 1993; Tanaka & Sengco, 1997). However, psychometric issues exist concerning the use of difference measures (see, e.g., Cronbach & Furby, 1970; Williams & Zimmerman, 1996). In addition, as yet there have been no investigations of the reliability or validity for the difference measures used in this study. Thus, for the relevant tasks, we do not concentrate solely on the difference measure between two conditions, but also use the performance in each condition as an indicator of the abilities related to face cognition.

The analyses of all reaction times (RTs) were based only on trials associated with correct responses. RTs were set to missing if they were less than 200 msec or were longer than 3.5 intraindividual standard deviations (*SD*s) above the individual's mean RT for a specific task. Participants' mean RTs for a particular task were defined as outliers and also set to missing if they were more than three *SD*s above the group mean RT. Missing data for RTs, as defined above, were totaled in each condition for each task. If, for a given participant, more than 40% of the data in a specific condition and task were missing, the individual's mean RT and accuracy for the specific condition of that task were set to missing. This was the case in 1.8% of all possible values.[2] If a given participant's data were set to missing for more than 5 of the 18 tasks, that participant

Table 1
Overview of the Tasks, Showing the Specific Affiliation With the Domain of Perceiving (P), Learning (L), or Recognizing Either Faces (R) or Expressed Emotions (E), the Task Type, Serial Position in the Study, Duration, and the Numbers of Trials and of Unique Facial Identities Used

Task	Name of Task	Domain	Speed/ Accuracy	Serial Position	Duration (min)	No. of Trials	No. of Faces
1	Simultaneous matching of morphs	P	Speed	18	2.3	30	30
2	Simultaneous matching of half-faces	P	Speed	16	5.2	60	30
3	Simultaneous matching of faces from different viewpoints	P	Speed	6	2.8	30	60
4	Face/nonface classification	P	Speed	8	10.4	240	120
5	Facial resemblance	P	Accuracy	3	9.2	48	32
6	Sequential matching of part–whole faces	P	Accuracy	9	5.8	60	30
7	Simultaneous matching of spatially manipulated faces	P	Accuracy	11	12.3	60	30
8	Delayed nonmatching to sample	L	Speed	4	4.3	30	60
9	Facial short-term memory	L	Accuracy	7	23.1	150	300
10	Acquisition curve	L	Accuracy	1	18.5	150	180
11	Cambridge Face Memory Test	L	Accuracy	13	15.0	72	52
12	Recognition speed of learned faces	R	Speed	10	20.0	32	32
13	Priming of learned faces	R	Speed	15	12.0	120	120*
14	Decay rate of learned faces	R	Accuracy	14	2.4	30	60*
15	Eyewitness testimony	R	Accuracy	5	3.2	30	60†
16	Facially expressed emotion decision	E	Speed	2	2.2	30	30
17	Emotional odd-man-out	E	Speed	17	3.1	30	90
18	Facially expressed emotion labeling	E	Accuracy	12	3.1	30	30

Note—Speed/Accuracy, predominant source of performance variability; No. of Faces, number of different facial identities. The familiar faces used in Tasks 13–15 were from the acquisition curve (*) or delayed nonmatching to sample (†).

was excluded from all analyses. On this basis, 2 participants were omitted. When participants were missing only a few data points, we replaced these values using the EM algorithm implemented in SPSS 12.0 (SPSS Inc., Chicago, IL). The MCAR test, following Little (1988), was not significant [$\chi^2(4355) = 4{,}377$, $p = .41$], indicating that the assumption of the randomness of missing values could not be rejected.

The idea of setting observations that were actually collected to missing values and to replace these data with estimates might seem awkward. The rationale for this procedure was to attempt to exclude all data from the analysis that probably reflect invalid observations. This might be the case if participants did not fully understand the instructions or did not succeed in figuring a way to optimally solve a specific task. Rather than completely excluding the data in such cases, which included many valid observations, from further analysis, it seemed more appropriate to eliminate the probably invalid observations and to replace them with adequate values. For the large amount of data and the small proportion of missing data in this study, this procedure had little or no apparent effect on the results we report.

In the following sections, all tasks are described individually. They are organized according to the process that the task was supposed to measure. At the beginning of each process section, shared methodological features for the group of tasks are described. For each task, relevant theoretical background is given prior to a description of its procedure. The performance indicators, as well as the descriptive and psychometric results, are briefly summarized. Apart from means, *SD*s, and standard errors (*SE*s) for RTs and accuracy rates, the internal consistency estimate of reliability (Cronbach's α) and ω, an indicator of factor saturation, are provided. The indicators used for computing α and ω were based on groups of trials, in order to maximize the number of cases incorporated. We computed ω via exploratory factor analysis, whereby a single latent factor was specified for each task on the basis of its group of indicators. The ω statistic represents the proportion of variance in the scale score that is accounted for by the latent variable that is common to all scale indicators; this is expressed as the scale's general factor saturation (McDonald, 1999; Zinbarg, Revelle, Yovel, & Li, 2005; Zinbarg, Yovel, Revelle, & McDonald, 2006). If a particular task was intended to study the experimental effects of differences between conditions, we also report the statistical test for the difference score as well as the effect size d, and end with a brief discussion of the existing literature.

FACE PERCEPTION TASKS

To assess perception of faces as purely as possible, these tasks were designed to minimize any unwanted influence from other abilities, such as memory or recognition. Thus, in the majority of the tasks, the face stimuli to be compared were presented simultaneously on the screen. In these simultaneous matching tasks, one face was positioned in the upper left quadrant of the screen and the other in the lower right, to minimize direct comparisons of individual features of the faces.

In the tasks with two conditions (Tasks 2, 6, and 7), the same pairs of stimuli in a trial were used in both conditions, in order to make the conditions as comparable as possible. For these stimulus pairs, restrictions were applied across conditions to pseudorandomly allocate (1) the position (i.e., top left or bottom right) of each stimulus within the pair, with the restriction that positions were different across conditions only 50% of the time; and (2) which condition was presented first for a particular stimulus pair, with the restriction that 50% of pairs were shown first in Condition 1. Finally, the sequence of trials was randomly generated, with the restrictions that all stimuli were shown once before any were repeated and that the minimal lag between the stimulus repetitions was 10 trials.

Task 1: Simultaneous Matching of Morphs

One of the most basic tasks in face perception is to determine whether two faces are the same or different. When two different faces are presented, the task is very simple if one face has, for instance, a relatively big nose; the decision in this case can be based on the comparison of a single feature. In order to minimize such feature-based processing, we employed the morphing procedure (Busey, 1998; Preminger, Sagi, & Tsodyks, 2007) using the Morpher 3.0 software (www.asahi-net.or.jp/~FX6M-FJMY/mop00e.html). Another benefit of the morphing procedure is that it allows for precise manipulations of difficulty through systematic manipulations of stimulus similarity. Morphing creates a new face (a *morph*) out of two different *parent* faces by combining and averaging the features of those faces. For a given pair of parent faces, a continuum of morphed faces can be derived that differ in the relative contributions of each parent face (e.g., 10% of parent A and 90% of parent B).

Procedure. In each trial, two nonidentical morphed faces, derived from the same parent faces, were presented; the task was to decide whether they were similar or dissimilar. Faces in the similar trials were closer to each other in the morphing continuum than were dissimilar faces. In the similar trials, Face 1 consisted of 50% of parent A and 50% of parent B, whereas Face 2 consisted of 30% A and 70% B. In the dissimilar trials, Face 1 consisted of 20% A and 80% B, whereas Face 2 consisted of 80% A and 20% B. Each parent face was used only once in each type of trial, and always in combination with a different randomly selected parent face, in order to prevent familiarization with the stimulus materials and to maintain the criterion for unfamiliar face perception. The face stimuli were taken from the Psychological Image Collection at Stirling (PICS, pics.psych.stir.ac.uk/). This test was intended to be a speed task.

Results. Average RTs and accuracy rates, as well as the αs and ωs for the parameter of interest (speed or accuracy), are shown in Table 2 for this task and all other perception tasks. RTs here were short, and average accuracy rates were very high. The results for α and ω both

Table 2
Descriptive Data for Reaction Times (in Milliseconds) and Accuracy Rates for All Face Perception Tasks, With Homogeneity Coefficients for the Parameter of Interest (Reaction Time for Tasks 1–4; Accuracy for Tasks 5–7)

	Reaction Time			Accuracy			
Condition	*M*	*SD*	*SE*	*M*	*SD*	*SE*	*α*/*ω*
Task 1: Simultaneous Matching of Morphs							
All conditions	1,543	397	63	.93	.06	.01	.84/.84
Similar (15)	1,472	479	39	.97	.05	.01	–
Dissimilar (15)	1,634	430	35	.89	.12	.01	–
Task 2: Simultaneous Matching of Half-Faces							
Aligned	1,703	533	85	.96	.06	.01	.88/.89
Nonaligned	1,748	530	84	.95	.06	.01	.82/.84
Task 3: Simultaneous Matching of Faces From Different Viewpoints							
All conditions	2,232	661	105	.92	.06	.01	.88/.88
Same face (15)	2,484	899	73	.89	.10	.01	–
Different faces (15)	2,019	638	52	.95	.07	.01	–
Task 4: Face/Nonface Classification							
Mean typical Caucasian	643	142	23	.98	.02	.00	.90/.91
Face typical Caucasian (30)	652	177	14	.98	.02	.01	–
Nonface typical Caucasian (30)	634	135	11	.98	.03	.01	–
Mean distinct Caucasian	643	135	22	.99	.02	.00	.91/.92
Face distinct Caucasian (30)	648	173	14	.99	.03	.01	–
Nonface distinct Caucasian (30)	640	125	10	.98	.02	.01	–
Mean Caucasian	622	127	20	.99	.02	.00	.90/.91
Face Caucasian (30)	626	160	13	.98	.03	.01	–
Nonface Caucasian (30)	620	116	9	.99	.02	.01	–
Mean Asian	651	127	20	.98	.02	.00	.88/.88
Face Asian (30)	660	158	13	.97	.03	.01	–
Nonface Asian (30)	642	125	10	.99	.02	.01	–
Task 5: Facial Resemblance							
All conditions	4,622	1,641	262	.72	.09	.01	.53/.54
Easy (16)	3,831	1,258	102	.83	.11	.01	–
Medium (16)	4,651	1,738	141	.75	.13	.01	–
Difficult (16)	5,672	2,506	203	.59	.13	.01	–
Task 6: Sequential Matching of Part–Whole Faces							
Part	1,513	329	53	.74	.11	.02	.58/.58
Whole	1,732	518	83	.71	.12	.02	.57/.58
Task 7: Simultaneous Matching of Spatially Manipulated Faces*							
Mean upright	5,430	2,311	369	.69	.11	.02	.55/.56
Upright eyes up or down (6)	4,666	2,495	208	.54	.26	.02	–
Upright eyes in or out (4)	5,228	3,369	299	.44	.30	.02	–
Upright mouth up or down (5)	3,854	1,734	142	.77	.24	.02	–
Mean inverted	6,092	2,653	423	.64	.11	.02	.43/.48
Inverted eyes up or down (6)	5,702	3,129	260	.51	.27	.02	–
Inverted eyes in or out (4)	5,317	2,873	244	.52	.28	.02	–
Inverted mouth up or down (5)	5,213	2,444	206	.60	.30	.02	–

Note—Numbers in parentheses report the number of trials per condition. *Subconditions for Task 7 are reported for trials with different faces only.

showed this to be a speed task with very good psychometric qualities.

Task 2: Simultaneous Matching of Half-Faces

Evidence supporting the role of holistic processing in face perception is derived from the so-called *part–whole effect* (see Task 6) and the *composite-face effect* (see, e.g., Hole, 1994; Young, Hellawell, & Hay, 1987). The latter refers to the phenomenon in which two complementary (i.e., upper and lower) half-faces from different people appear to fuse into a new face, in which the internal features are strongly integrated, when the halves are aligned. This effect makes it difficult to parse the face into isolated features—impeding, for example, recognition or naming of one half-face when the halves are aligned, rather than offset (i.e., nonaligned; see Hole, 1994; Young et al., 1987). The advantage of the nonaligned condition is found only when the stimuli are presented upright as opposed to upside down, supporting the notion that holistic and configurational types of processing are recruited in normal (i.e., upright) perceptual processing of face stimuli but are impaired in the inverted condition.

Procedure. Faces were divided horizontally (approximately halfway down the nose) into upper and lower halves

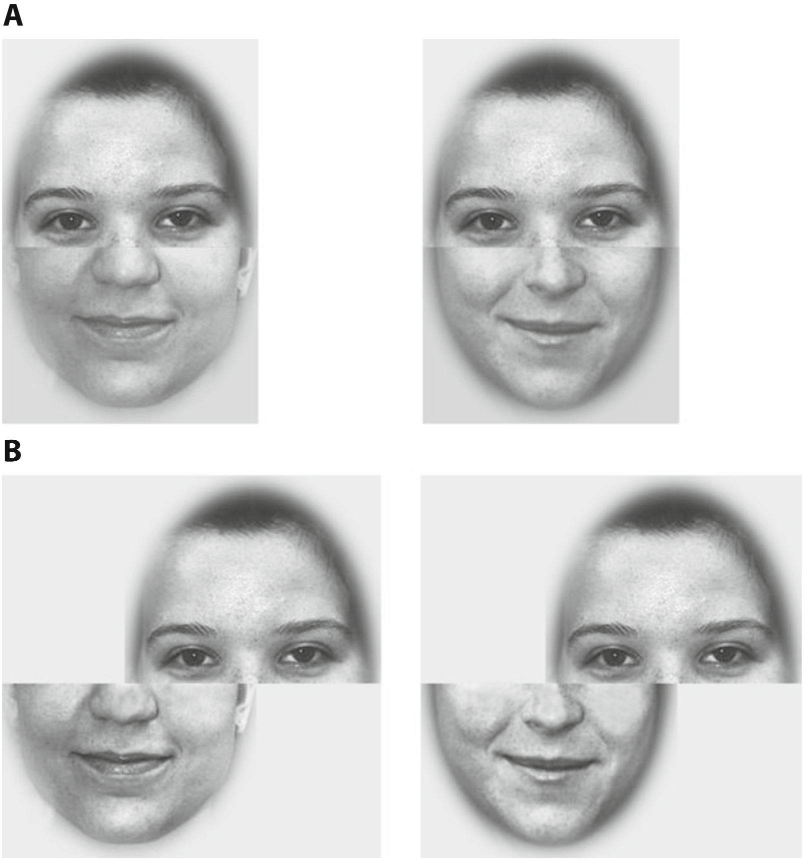

Figure 1. Example of a stimulus set for Task 2 (simultaneous matching of half-faces). Panel A shows an example of the aligned and panel B of the nonaligned condition. The faces are from Endl et al. (1998).

and coupled with a complementary half from another face, forming a new face. In each trial, two new composite faces were presented, and the task was to decide whether the top halves were the same or different; the lower halves were always different. Nonaligned stimuli were joined so that the left or right edge of the top half-face was positioned at the nose of the bottom face (see Figure 1). Within each condition, each original face was used twice as a top half (either in one similar trial or two dissimilar trials) and twice as a lower half, always in combination with an upper half from a different face. The same stimulus pairs were used across the conditions. The face stimuli were taken from the Extended M2VTS Database (XM2VTSDB; Messer, Matas, Kittler, Luettin, & Maitre, 1999). This test was designed as a speed task.

Results and Discussion. Table 2 summarizes the main results. Mean RTs in the aligned and nonaligned conditions were short, and accuracies were very high, as expected for a speed task. Both α and ω showed good psychometric qualities.

Contrary to the previous literature (e.g., Hole, 1994; Young et al., 1987), responses to aligned faces were 45 msec faster and also more accurate than those to nonaligned faces. Differences in RTs and accuracy (aligned − nonaligned) were statistically significant [*t*s(150) = 2.1 and 1.9, *p*s < .05] but small (*d*s = 0.08 and 0.17). The α for the difference in RTs was less than zero (ω could not be computed), and it was .23 (ω = .29) for differences in accuracy rates. Because, unlike in much of the literature, we did not obtain a composite-face effect, and because the effect we did obtain had poor psychometric qualities, we do not recommend using the difference between the conditions of this task as a measure of individual differences. Nevertheless, there can be little doubt that each of the conditions did measure processes underlying face perception with good psychometric qualities. Therefore, these conditions seem to be meaningful measures of individual differences in face perception.

Task 3: Simultaneous Matching of Faces From Different Viewpoints

Successful encoding of a face requires not only the extraction of pictorial codes (e.g., pose) but also the extraction and retention of expression- and viewpoint-independent information about the specific featural and configurational aspects of the face. Changing the viewpoint from which a face is depicted from study to test is thus used in face tasks to tap the extent to which the face has been structurally encoded. For both matching and recognition of unfamiliar faces, there is a disadvantage apparent both in accuracy and RTs either when the two faces are shown from different viewpoints or when the viewpoint changes between the initial and test presentations (see, e.g., Bruce, 1982; Newell, Chiroro, & Valentine, 1999).

Procedure. In this task, two faces were shown simultaneously, one in frontal view and the other in three-quarter view. The task was to indicate whether the faces depicted the same or different persons, and each trial used novel faces. The three-quarter views were always of the left side of the face. The stimuli were taken from Schacht, Werheid, and Sommer (2008). The task was intended to focus on speed.

Results. The results are summarized in Table 2. Mean RTs were a little higher than in the other speed tasks, and accuracy was as expected for a speed task. Both α and ω were high, showing this task to be a good speed task.

Task 4: Face/Nonface Classification

In face classification tasks, participants distinguish between faces and nonfaces. Usually in these tasks, typical faces and own-race faces are classified faster than distinctive or other-race faces (see, e.g., Valentine, 1991; Valentine & Bruce, 1986)—that is, face classification shows a *distinctiveness effect* and *own-race bias*, respectively. A theoretical explanation for these phenomena is given by the concept of the *face space* (Valentine, 1991). Face space is postulated to be multidimensional, defined by dimensions that serve to encode and discriminate faces, with the origin of the space representing the central tendencies of the dimensions. Within this framework, typical faces are clustered near the center of the space, with faces becoming more sparsely distributed as one moves toward the poles of the dimensions, in which distinctive faces are located. It is thought to be easier to classify a typical unfamiliar face as being a face because a greater number of typical familiar faces are located near the center than toward the poles of the dimensions. Thus, if a typical but unfamiliar face is viewed, it resembles many more of the faces that are already represented in the face space, whereas distinctive unfamiliar faces will have fewer matches. Typical faces are also thought to be much more similar to a prototypical face, and therefore easier to classify (Valentine & Bruce, 1986). In addition, Valentine (1991) suggested that the face space is finely tuned to own-race faces and that other-race faces are poorly represented. As a result, other-race faces not only would be represented farther away from the center but also would be more tightly clustered, because they are less individuated than own-race faces (Valentine, 1991), and hence are classified more slowly as faces.

Procedure. In each trial, either a face or a scrambled face was presented; participants then had to decide whether the stimulus was a face or not. To make a scrambled face, a 2 × 2 grid was fitted into the inner area of the face, affecting only the internal features, leaving the hairline as well as the face line intact. The four squares were randomly reshuffled to form a new combination. The squares were delineated by dark gray lines in both the scrambled and intact faces. Each intact face was also used as a scrambled face.

For the distinctiveness effect,[3] 120 trials were presented, with 30 typical and 30 distinctive faces; another 120 trials probed the own-race bias, with 30 Caucasian and 30 Asian faces. For each condition, 60 scrambled faces were used. All trials were randomly interspersed. To examine the own-race bias, responses to Caucasian and Asian faces were compared. The face stimuli were obtained from the CAS-PEAL database (Wen et al., 2004) and the Color FERET database (Phillips, Moon, Rizvi, & Rauss, 2000). This test was intended as a speed task, and the mean correct RTs in the typical, distinct, Asian, and Caucasian conditions were all of focal interest.

Results and Discussion. Table 2 summarizes the results. RTs for typical and distinct Caucasian faces were the same. Asian faces took 29 msec longer to classify than Caucasian faces.

Contrary to the literature (e.g., Valentine, 1991; Valentine & Bruce, 1986), we found no difference between the RTs or accuracy rates for distinct and typical Caucasian faces [ts(150) = 0.02 and 1.5, ps > .10]. For this distinctiveness effect, the αs for RTs and accuracy, respectively, were .14 and .18, and the ωs were .26 and .35. A possible reason for the absence of this effect could be that stimulus selection on the basis of distinctiveness ratings was not as efficient as in previous studies. However, using the same rating scale and procedure, Valentine (1991) found a distinctiveness effect in a classification task for typical and distinctive faces that showed mean ratings comparable to those in the present study. Because we failed to obtain a distinctiveness effect and because the difference between the experimental conditions showed poor psychometric qualities, the distinctiveness effect as measured in this task cannot be recommended as a measure of individual differences. However, each condition on its own did show good psychometric qualities. Therefore, performance in these conditions seems to be a meaningful measure of individual differences in face perception.

In line with the literature (e.g., Valentine, 1991; Valentine & Endo, 1992), we found that Caucasian faces were classified as faces significantly more quickly and accurately than Asian faces [ts(150) = 6.8 and 3.1, ps < .01]. The effect size for the difference in RTs was small (d = 0.22), and that for the difference in accuracy was moderate (d = 0.35). For the own-race bias, the statistics for RTs and accuracy, respectively, were α = .33 and .08, and ω = .45 and .17.

Task 5: Facial Resemblance

This task was inspired by the CFPT (Dingle et al., 2005), which was proposed as a test of perceptual face cognition that minimizes reliance on feature matching by using the morphing method.

In a pilot study, we used the same experimental settings and procedure as Dingle et al. (2005) and obtained results that were just above guessing probability. Thus, the task was too difficult to allow for an adequate assessment of the underlying ability in a broadly varying sample. The procedure was therefore adjusted as described below to facilitate performance of the task. We do not expect that these modifications would have changed what the task measures.

Procedure. In each trial, three faces were shown: One original target face, depicted in three-quarter view, was presented centered above two morphed faces, depicted in frontal view. Participants had to decide which of the morphs most resembled the target face. The morphed faces were derived from the morphed continuum of the original target face with a second parent face. For each target face, three levels of difficulty were created, generated by the distance in the morphing continuum between the two depicted morphs. In the easiest condition, Morph 1 consisted of 90% target and 10% other, and Morph 2 of 40% target and 60% other. In the medium-difficulty condition, the target face contributed 80% to Morph 1 and 50% to Morph 2. In the most difficult condition, Morph 1 consisted of 70% and Morph 2 of 60% of the target face. The stimuli were formed from 16 target faces, morphed with 16 other faces. Each target face was combined with only one other, and the morphs from each pair were represented in all three difficulty levels. The face stimuli were taken from Schacht et al. (2008). This test was intended as an accuracy task.

Results and Discussion. The results are summarized in Table 2. Mean accuracy was halfway between guessing and perfect accuracy. Both α and ω indicated that this task did not meet strict criteria for unidimensionality and homogeneity. There was no ceiling effect, as indicated by a Kolmogorov–Smirnoff test for normal distribution of the data [K–$S(151) = 0.93, p > .35$]. The guessing probability of 50% might be responsible for the psychometric issues. Increasing the number of distractors and/or increasing the number of trials should improve the psychometric properties in future versions of this task.

Task 6: Sequential Matching of Part–Whole Faces

The notion of holistic face cognition suggests that faces are normally perceived primarily as undifferentiated wholes with no (or little) independent representation of individual internal features. The *part–whole recognition effect* refers to the finding that a particular facial feature (e.g., a nose) is recognized more easily in the context of the face to which the feature belongs than when tested in isolation (Tanaka & Farah, 1993; Tanaka & Sengco, 1997). This advantage in the "whole" context is not found for other objects, such as houses, and is eradicated by inverting the face (Tanaka & Sengco, 1997), supporting the notion that holistic processing is special to the processing of upright faces. On the basis of such evidence, Tanaka and colleagues suggested that in normal face cognition, a particular facial feature is encoded in combination with its spatial relations with the other facial features.

Procedure. A target face (see Figure 2 for an example) was presented for 1,000 msec, followed by a mask for 200 msec. The mask consisted of three *X*s, centrally displayed, and covered the area of the target face. Next, a facial feature from the target (e.g., its nose) was presented along with the same feature from a different face. This facial feature was displayed in either of two conditions: in the context of the whole target face (e.g., the target's nose vs. another nose presented within the target face) or as isolated features (e.g., simply the target's nose vs. another nose). The task was to discern either which face was the target (in the whole condition) or which feature belonged to the target (in the part condition). The central facial features of eyes, nose, and mouth were tested separately and in equal proportions. To form the stimuli, each face was used once as a target and once as a distractor, but in each case was paired with a different face. The same stimulus pairs were used across the conditions. Face stimuli were

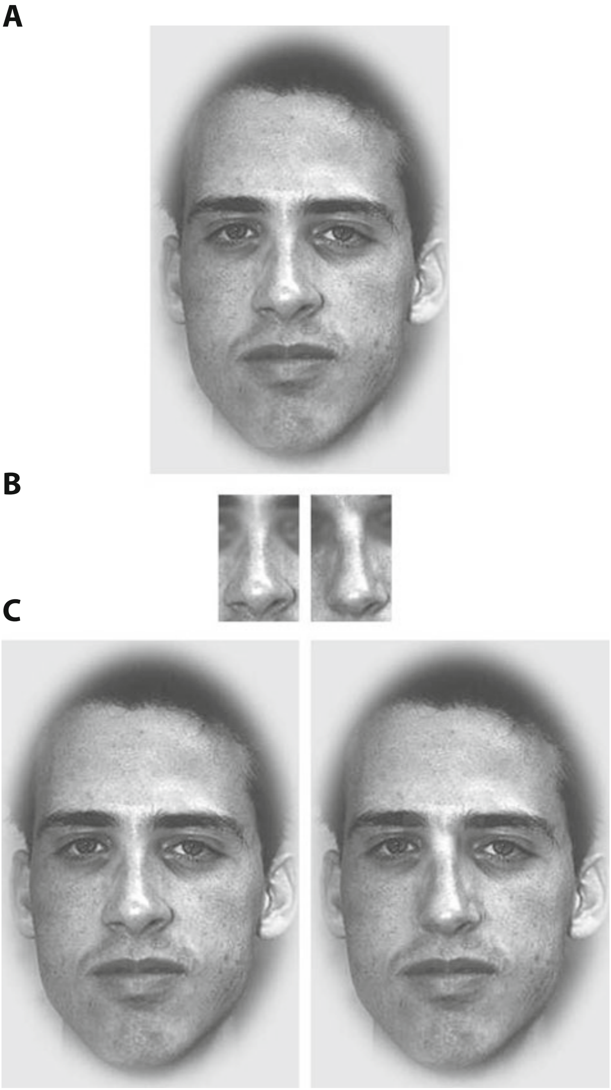

Figure 2. Example of a stimulus set for Task 6 (sequential matching of part–whole faces). Panel A shows the target face. Feature testing (here, with the nose) occurs either in isolation (panel B) or in the context of the target face (panel C; original nose on the left and different nose on the right). The faces are from Endl et al. (1998).

taken from the PICS database (see Task 1 above). This task was designed to focus on accuracy.

Results and Discussion. The results are summarized in Table 2. Accuracy was halfway between guessing and perfect accuracy. Both α and ω showed that the task would not allow for sufficient precision if used as a standalone indicator. On the other hand, viewed in conjunction with the number of trials (30 per condition) and the guessing probability, there was no psychometric problem with either indicator. Mean RTs were rather short and might add information about individual differences in performance on the task, as supported by αs of .83 and .79 and ωs of .83 and .80 for the part and whole conditions, respectively.

In contrast to the literature (e.g., Tanaka & Farah, 1993; Tanaka & Sengco, 1997), we found that judgments in the part condition were more accurate and faster than those in the whole condition [ts(150) = 2.8 and 6.9, ps < .01]. The effect size was small ($d = 0.25$) for accuracy rates but large ($d = 0.51$) for RTs. Differences (whole − part) in accuracy and RTs showed small to moderate αs (.22 and .56, respectively) and ωs (.26 and .57). Because we failed to obtain a face superiority effect and because the differences between the experimental conditions showed poor psychometric qualities, the effect obtained here cannot be recommended as a measure of individual differences in the size of face superiority. However, because each condition on its own showed good psychometric qualities, performance in these conditions was a meaningful measure of individual differences in face perception in a more general sense.

Task 7: Simultaneous Matching of Spatially Manipulated Faces

Much evidence for the importance of configurational information in face perception comes from the so-called *face inversion effect* and from the effect on performance of manipulating the spatial relations between facial features. Many studies have shown that turning a face upside down makes face perception and recognition slower and less accurate. Importantly, this inversion effect is disproportionately greater for faces than for other objects (see, e.g., Yin, 1969; see Valentine, 1988, and Searcy & Bartlett, 1996, for reviews). Another important consequence of inverting faces is that spatial (or configurational) displacements of features are harder to detect, as compared with an upright condition. This effect is much greater when only local features are changed (e.g., Freire, Lee, & Symons, 2000). Interestingly, recent studies have shown that inversion affects spacing discrimination as much as it does local-feature discrimination (e.g., Yovel & Duchaine, 2006). Taken together, the face inversion effect is thought to result mainly from a disruption to the processing of configurational information, and is therefore suggested to be diagnostic for configurational processing (Maurer, Le Grand, & Mondloch, 2002).

Procedure. In this task, participants had to indicate whether two simultaneously presented faces were the same or different. The faces were always derived from the same picture, but in the *different* condition, one spatial relationship between features was altered from the original. The spatial manipulations varied in extent, thereby manipulating difficulty, and were as follows: (1) moving the eyes up or down by 9, 11, or 13 pixels; (2) moving the eyes in or out by 10 or 12 pixels in total; and (3) moving the mouth up or down by 7, 9, or 11 pixels—thus changing either the eyes–nose or the mouth–nose relation (see Figure 3 for an example). For the eyes up/down manipulation, the area containing the eyes, eyebrows, and bridge of the nose was shifted vertically. The faces did not depart from the range of normal variation after the spatial changes; this was partly achieved by using only typical faces as the stimuli to be manipulated. The same stimuli were used across conditions. The face stimuli were taken from PICS (see Task 1). This test was intended as an accuracy task.

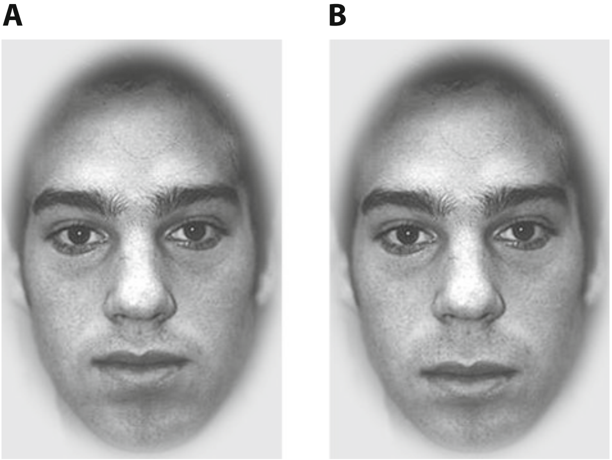

Figure 3. Example of a stimulus set for Task 7 (simultaneous matching of spatially manipulated faces). Panel A shows the original face. In panel B, the mouth–nose relation is altered by moving the mouth down by 11 pixels. The faces are from Endl et al. (1998).

Results and Discussion. Mean accuracy and RTs are shown in Table 2. Both α and ω indicated that performance was not highly reliable (Table 2). Still, in the context of a multivariate battery, both conditions might contribute valuable information. If the task were to be used as a standalone indicator, more trials would be necessary.

In line with the literature (e.g., Searcy & Bartlett, 1996; Valentine, 1988), judgments in the upright condition were more accurate and faster than those in the inverted condition [ts(150) = 5.5 and 6.7, ps < .001]. The effect size was medium ($d = 0.67$) for accuracy and small ($d = 0.27$) for RTs. The differences (upright − inverted) in accuracy and RTs showed small αs (.17 and .19, respectively) and ωs (.24 and .16).

FACE LEARNING TASKS

Our approach to separate learning and recognition processes was to manipulate the time between the initial presentation and the recognition test of the stimulus. When this time is short, the contribution of long-term memory (as typically defined) to task performance is minimized, and the contributions of learning processes and short-term

Table 3
Descriptive Data for Reaction Times (in Milliseconds) and Accuracy Rates for All Face Learning Tasks, With Homogeneity Coefficients for the Parameter of Interest (Reaction Time for Task 8; Accuracy for Tasks 9–11)

	Reaction Time			Accuracy			
Condition	*M*	*SD*	*SE*	*M*	*SD*	*SE*	α/ω
Task 8: Delayed Nonmatching to Sample							
All conditions	1,116	247	40	.97	.04	.01	.90/.90
Task 9: Facial Short-Term Memory							
Nonselected	1,645	329	53	.90	.09	.02	.63/.64
Typical Caucasian	1,955	456	73	.82	.12	.02	.68/.68
Distinct Caucasian	1,680	376	60	.89	.10	.02	.70/.70
Caucasian	1,707	372	59	.90	.08	.01	.50/.52
Asian	2,079	527	84	.87	.10	.02	.63/.65
Task 10: Acquisition Curve							
All blocks	1,627	370	59	.89	.07	.01	.89/.90
Block 1	2,013	596	95	.85	.01	.01	.54/.56
Block 2	1,784	524	83	.88	.01	.01	.63/.64
Block 3	1,601	415	66	.91	.01	.01	.69/.70
Block 4	1,479	335	54	.90	.01	.01	.64/.65
Block 5	1,310	249	40	.93	.01	.01	.64/.65
Task 11: Cambridge Face Memory Task							
Block 1 (18)	2,761	633	101	.98	.05	.01	.37/.38
Block 2 (30)	4,370	1,240	198	.69	.15	.02	.81/.81
Block 3 (24)	3,925	1,002	160	.57	.16	.03	.68/.70

Note—Numbers in parentheses report the number of trials per condition.

memory are higher. For all tasks in this section, a relatively short delay of at most 4 sec between learning and recognition, consisting of a mask (500 msec) and a blank screen (3,500 msec), was instantiated.

Task 8: Delayed Nonmatching to Sample

Delayed nonmatching-to-sample (DNMS) and delayed matching-to-sample (DMS) tasks are widely used to investigate visual short-term memory. Both tasks require a participant to hold a visual stimulus "online" over a delay interval before responding to a choice. A trial-unique target stimulus is presented in both tasks, followed by a blank-screen delay interval, after which the target and a novel stimulus are presented. The participant must select the novel stimulus in DNMS and the familiar stimulus in DMS tasks. Whereas human infants and monkeys show an instinctive preference for the novel stimuli (A. Diamond, 1991; Gaffan, Gaffan, & Harrison, 1984), human adults have a strong bias to match—that is, to identify the familiar target stimulus—at the choice stage (see, e.g., Aggleton, Nicol, Huston, & Fairbairn, 1988; Elliott & Dolan, 1999). Thus, as compared with DMS tasks, the DNMS task requires an additional process of response inhibition. Here, we conducted a unique-trial DNMS task that required the encoding of an unknown face and—after a delay—the distinction of the learned face from a new one.

Procedure. The target face was shown for 1,000 msec, followed by the delay, after which the target face was presented together with a new face. Participants had to indicate the novel face. The face stimuli were taken from the Color FERET database (Phillips et al., 2000). This test was designed as a speed task.

Results. Average RTs and accuracy rates, as well as αs and ωs for the parameters of interest (speed or accuracy), for this and all other learning tasks are shown in Table 3. The average RT in Task 8 was very short, and accuracy was very high. Both α and ω indicated good psychometric qualities for a speed task.

Task 9: Facial Short-Term Memory

Visual short-term memory tasks usually require the learning of a set of elements that have to be recalled or recognized after a short delay. There are only a few studies investigating short-term memory for facial material in which more than a single face has to be learned at a given time and then retrieved (e.g., Bruyer & Vanberten, 1998; Hanley, Young, & Pearson, 1991); all of these have required sequential learning of the faces and later the recall of their serial positions. Because we wanted to measure learning over a short period of time, we designed a task in which a set of faces had to be learned simultaneously and recognized after a short delay.

Integrated within this task were stimulus specificities enabling the investigation of the distinctiveness effect and the own-race bias. In tasks requiring learning, distinctive as well as own-race faces are learned better than typical or other-race faces, respectively (see, e.g., Valentine, 1991; Valentine & Endo, 1992). According to the face space hypothesis (Valentine, 1991), typical faces are clustered around the center of the face space. Learning of new representations is thought to be more difficult for typical faces, as compared with distinctive faces, because they are more likely to be confused with the many other typical faces in their neighborhood. Other-race faces are considered more difficult to learn than own-race faces because their representations in the face space are suggested to be more tightly clustered and less individuated.

Procedure. The task comprised 15 blocks, each of which began with the simultaneous presentation of a set of 10 different and always novel target faces. Participants had 45 sec to learn the set. After a delay, each target face was presented together with a new face. Participants had to indicate whether the previously seen face was on the left or the right. After testing all target faces for recognition, the next block started. Six blocks of this task investigated the distinctiveness effect with Caucasian faces, and 6 others probed the own-race bias with Caucasian and Asian faces. In the remaining 3 blocks, only nonselected[4] Caucasian faces were used in order to investigate general short-term memory for faces, independent of the distinctiveness effect and own-race bias. The face stimuli were taken from XM2VTSDB (Messer et al., 1999) and CAS-PEAL (Wen et al., 2004). This test was devised as an accuracy task.

Results and Discussion. Table 3 summarizes the results for this task. Overall, the accuracies show adequate but lower than desirable homogeneity, again possibly attributable to the limited test length and the high guessing probability of 50%. The RTs might account for some individual variation, as supported by both α and ω ranging between .77 and .85.

In line with the literature (e.g., Valentine, 1991; Valentine & Endo, 1992), distinctive faces were learned more easily than typical faces [$t(150) = 10.0$, $p < .001$]. The effect size for the accuracy rates was medium ($d = 0.65$), and the associated α (.21) and ω (.25) were small. Also consistent with the literature (e.g., Valentine & Endo, 1992), own-race faces were learned more easily than other-race faces [$t(150) = 5.3$, $p < .001$]. The effect size for the accuracy rates was medium ($d = 0.41$), and α for the difference in accuracy was small (.22; $\omega = .28$).

Task 10: Acquisition Curve

In everyday life, face learning is rarely a one-trial process, because we see most relevant faces repeatedly. The number of repetitions necessary until a face is successfully recognized can be considered to indicate how well a person learns. In this task, we aimed to measure the learning success for a large number of faces over several runs.

Procedure. In the study phase, 30 faces, randomly arranged in six rows, were presented on the screen for 2 min. Because all faces did not fit within the dimensions of the screen, participants were instructed to scroll up and down through all of the faces during learning. Proctors reminded the participants, after 1 min, of the faces in the bottom rows. The test phase began 4 sec after the screen was cleared and consisted of five runs. Within each run, each studied target face was presented on a trial-by-trial basis, alongside a completely new face. Participants had to indicate which was the target face. Immediately after their response, the target face was highlighted by a green frame, regardless of the accuracy of the response. After a blank screen for 500 msec, the next trial started. The position of the target (left or right) was randomly assigned, with the restriction that half of the targets changed their position from run to run, whereas the other half were presented in the same position as in the preceding run. The face stimuli were taken from the Color FERET database (Phillips et al., 2000) and from Schacht et al. (2008). This task was designed to focus on accuracy.

Results and Discussion. Results for this task are summarized in Table 3. Correlations across blocks for this task, together with Task 14, which represents the long-term recognition test of the faces learned in this task (see below), are shown in Table 4. The αs show only adequate homogeneity, which again could be attributable to the 50% guessing probability. Although mean performance was already high in Block 1, a steady increase was seen across the following blocks. The mean accuracy in Block 5 was skewed but not yet at ceiling. The variance in performance that remained in Block 5 was substantial enough to allow for high correlations with performance in the previous blocks. The correlations between blocks generally approached, and sometimes exceeded, the estimates for internal consistency, and thus were as high as could be expected. Interestingly, there was a decline in RTs across blocks, visible also in the RT *SE*s and *SD*s. The RTs might potentially provide additional information about individual differences not provided when considering the accuracies alone; this is supported by αs ranging between .85 and .89, as well as by ωs between .85 and .90.

Table 4
Correlations and α Coefficients for Accuracy Rates Across Blocks for Task 10 (Acquisition Curve) and Task 14 (Decay Rate)

	Block 1	Block 2	Block 3	Block 4	Block 5	Decay Rate
Block 1	.54	–	–	–	–	–
Block 2	.58	.63	–	–	–	–
Block 3	.55	.68	.69	–	–	–
Block 4	.56	.59	.66	.64	–	–
Block 5	.65	.63	.70	.69	.64	–
Decay rate	.57	.54	.58	.60	.61	.58

Note—The α coefficients are on the main diagonal.

Task 11: Cambridge Face Memory Task

The CFMT (Duchaine & Nakayama, 2006) is a test of face learning and short-term recognition, similar to those used in the tasks designed here (e.g., Tasks 9 and 10). The main difference between the stimuli here and those in the other tasks is that in the CFMT, even the hairline is masked; the task comprises three blocks of varying difficulty, primarily instantiated by varying degrees of stimulus degradation. The CFMT presents the same items in both an upright and an inverted version. Duchaine and Nakayama showed that (1) normal adults performed poorly on the inverted version and (2) prosopagnosic patients performed poorly even on the upright version. The CFMT is thus considered a test of the skills specific to face cognition.

Procedure. The procedure of the CFMT is described in detail in Duchaine and Nakayama (2006); all of the stimuli are male faces. We used the upright version of the CFMT only. This test was intended as an accuracy task.

Results and Discussion. The results are summarized in Table 3. Accuracy on the CFMT was very high in the first block, but Blocks 2 and 3 were substantially harder. RTs were rather long. The αs and ωs for the accuracies of the three blocks were acceptable, and the overall α for the test was .83.

FACE RECOGNITION TASKS

In order to emphasize recognition over learning processes, all tasks in this section had a minimum delay of more than 4 min between learning and the corresponding recognition test. Because of the disadvantages of using preexperimentally familiar faces, the faces for the recognition tasks were learned in a session preceding the actual recognition test (Tasks 12, 13, and 14) or had been distractor stimuli in a previous task, and thus had not been explicitly learned (Task 15). Except in Task 15, participants had to complete other tasks between learning and recognition. Please note that in these tasks we did not intend to measure familiar face recognition, but instead were interested in the cognitive processes underlying the ability to recognize learned/familiar faces.

Task 12: Recognition Speed of Learned Faces

This task follows typical assessment procedures for recognition memory—for example, Warrington's (1984) Recognition Memory Test. Participants learned a number of faces that were subsequently tested for recognition. In

Table 5
Descriptive Data for Reaction Times (in Milliseconds) and Accuracy Rates for All Face Recognition Tasks, With Homogeneity Coefficients for the Parameter of Interest (Reaction Time for Tasks 12 and 13; Accuracy for Tasks 14 and 15)

	Reaction Time			Accuracy			
Condition	*M*	*SD*	*SE*	*M*	*SD*	*SE*	*α*/*ω*
Task 12: Recognition Speed of Learned Faces							
All conditions	1,380	289	46	.89	.09	.01	.75/.75
Learned face (16)	1,359	335	27	.86	.13	.01	–
Unfamiliar face (16)	1,402	336	27	.91	.10	.01	–
Task 13: Priming of Learned Faces							
Primed learned	850	268	43	.87	.10	.02	.74/.75
Unprimed learned	1,144	223	36	.84	.11	.02	.75/.75
Primed unfamiliar	1,171	367	59	.81	.13	.02	.83/.84
Unprimed unfamiliar	1,265	321	51	.88	.09	.02	.71/.73
Task 14: Decay Rate							
All conditions	1,913	416	66	.86	.01	.01	.58/.59
Task 15: Eyewitness Testimony							
All conditions	2,433	701	112	.65	.11	.02	.52/.53

Note—Numbers in parentheses report the number of trials per condition.

order to increase the demands on memory, we included a delay of at least 4 min between learning and recognition.

Procedure. The task was composed of four parts, each consisting of a learning phase followed by a delay of an average of 4.8 min, during which participants completed a different task, and the recognition test. In each learning phase, four faces were shown simultaneously for 30 sec and were to be memorized. During the delay period, participants completed four items of Raven's Advanced Progressive Matrices (Raven, Raven, & Court, 1998). At test, four learned and four completely unfamiliar faces were shown one at a time in a pseudorandomized order. For each face, participants indicated whether it had been presented in the learning session. The faces were taken from Carbon and Leder (2007). This test was a speed task.

Results and Discussion. Average RTs and accuracy rates, as well as αs and ωs for the parameter of interest (speed or accuracy), for this and all other recognition tasks are shown in Table 5. Mean RTs in this task were short and accuracies high. The accuracy distribution was skewed but not at ceiling. Both α and ω for the RTs were acceptable, but not very high as compared with the results for other speed indicators. Apparently, there was also a considerable amount of systematic variance in the accuracies, as indicated by the α of .69 and the ω of .70.

Task 13: Priming of Learned Faces

In memory research, *priming* refers to the facilitation or inhibition of a response to a given item due to a preceding prime stimulus. In familiarity decision tasks, face recognition is facilitated by the previous presentation of the same face but not by earlier presentation of the person's name, indicating that priming does not act at semantic stages in face cognition (see, e.g., Bruce & Valentine, 1985). The priming effect is larger for familiar than for unfamiliar faces, and for familiar faces it is also present when a different portrait of the same person is used as the prime. This suggests that priming of faces relies not only on pictorial codes but, more importantly, on view-independent representations of the facial structures stored in memory (e.g., Bruce & Valentine, 1985). Repetition priming of familiar faces has therefore been proposed to be mediated by residual activation of the FRU from the presentation of the prime face.

Procedure. Each trial began with a black fixation cross displayed for 200 msec, followed first by a prime face presented for 1 sec, then by a black fixation circle. After 1,300 msec, the circle was replaced by a target stimulus. Participants decided whether a target face, preceded by its prime face, was familiar (i.e., learned) or unfamiliar. The primes were to be ignored. The familiar faces in this task were the 30 faces that had been learned over 2.5 h earlier, during Task 10. A prime could be either the same stimulus (primed condition) or an unrelated stimulus (unprimed condition). In the unprimed condition, learned faces were used as primes for unfamiliar targets and unfamiliar faces as primes for learned targets. The factors of priming and familiarity were orthogonal, and levels of both factors were equiprobable. The stimuli were taken from the Color FERET database (Phillips et al., 2000) and from Schacht et al. (2008). The test was designed as a speed task.

Results and Discussion. The descriptive data do not unequivocally support the priming task as an indicator of the speed of face cognition (Table 5). Although, as indicated by the mean RTs, the task was not very complex, the accuracies were surprisingly low. Once again, it is plausible that the proportions of correct responses contained systematic individual differences. Indeed, αs for the accuracies ranged between .71 and .83, and ωs between .73 and .84, similar in magnitude to those for the RTs.

In line with the literature (e.g., Bruce & Valentine, 1985), RTs for primed stimuli were faster than those for unprimed stimuli in both the learned and unfamiliar conditions [$ts(150) = 19.1$ and 4.9, $ps < .001$]. The effect size was large ($d = 1.2$) for learned faces but small ($d = 0.27$) for unfamiliar faces. Priming effects (unprimed − primed) in RTs showed small αs of .35 and .40 ($\omega = .36$ and .41) for learned and unfamiliar faces, respectively.

Task 14: Decay Rate

Everything learned and represented in memory may be forgotten or become inaccessible. Therefore, the accessibility of previously learned faces after a longer delay may be an important indicator for face recognition. A necessary prerequisite for assessing such capabilities is a well-established and uniform memory trace across participants, which can be realized either by controlling the frequency of repeated stimulus presentations during learning or by controlling the level of accuracy in recognition of the stimuli at the end of the learning phase. For practical reasons, we controlled the frequency of presentations.

Procedure. The 30 faces that had been learned during Task 10 were tested for recognition after about 2.5 h. In each trial, two faces were shown to the left and right on the screen; one had been previously learned, and the other was new, and participants had to indicate the previously learned face. The face stimuli were taken from the Color FERET database (Phillips et al., 2000) and from Schacht et al. (2008). This test was designed as an accuracy task.

Results and Discussion. The main results are summarized in Table 5. There was a substantial decrease in performance from Block 5 of Task 10 to Block 6 (the present task), our operational definition of *decay rate*. Performance in the present task essentially declined to the level of Block 1—allowing for individual differences in performance decrement. The correlations between decay rate and the five blocks of Task 10 were very constant, around $r \approx .60$ (Table 4). Here, α was as high as in Task 10. Hence, the present data support the use of experimentally learned rather than preexperimentally familiar faces in recognition tasks as a possible indicator of individual differences in face recognition.

Task 15: Eyewitness Testimony

Eyewitness testimony refers to long-term retention and recognition of an event after a single exposure, either with an explicit learning intention or with implicit learning in circumstances in which the event was not relevant. Here we used implicit learning. The targets were distractor faces from the immediately preceding (in the test sequence) Task 8, during which no instruction about subsequent recognition testing had been given.

Procedure. In each trial, two faces were displayed side by side on the screen. Participants had to indicate the face that they had seen about 5 min before, in the preceding Task 8. The face stimuli were taken from the Color FERET database (Phillips et al., 2000). This test was designed as an accuracy task.

Results and Discussion. The results are summarized in Table 5. Mean accuracy in this task was not very high, and α and ω were rather low. We partly attribute this result to the small number of trials and the high guessing probability of 50%. Furthermore, this task was made especially hard to accomplish by requiring participants to incidentally learn faces and recognize them later. It can be argued that more than one ability is necessary for such a task. As a standalone measure, this test would not be acceptable psychometrically. In concert with a larger variety of additional measures, however, it can still be considered to contribute useful information, as an accuracy measure, for an overall assessment of the abilities related to face cognition.

RECOGNITION TASKS FOR FACIALLY EXPRESSED EMOTIONS

The following tasks were designed to assess the recognition of emotion. Because most models of face recognition consider this process to be independent of identity recognition, it was assumed that the ability to recognize and categorize emotional facial expressions might be separate from face learning and memory.

Table 6
Descriptive Data for Reaction Times (in Milliseconds) and Accuracy Rates for All Facially Expressed Emotion Recognition Tasks, With Homogeneity Coefficients for the Parameter of Interest (Reaction Time for Tasks 16 and 17; Accuracy for Task 18)

	Reaction Time			Accuracy			
Condition	*M*	*SD*	*SE*	*M*	*SD*	*SE*	α/ω
Task 16: Facially Expressed Emotion Decision							
All conditions	783	215	34	.96	.05	.01	.89/.90
Happiness	658	169	27	.98	.03	.01	–
Anger	922	304	49	.93	.09	.02	–
Task 17: Emotional Odd-Man-Out							
All conditions	2,623	622	99	.88	.07	.01	.76/.77
Happiness	2,934	1,045	167	.88	.15	.02	–
Sadness	2,171	674	108	.92	.12	.02	–
Anger	2,549	867	138	.96	.10	.01	–
Fear	2,898	967	154	.81	.13	.02	–
Surprise	2,298	663	106	.95	.11	.02	–
Disgust	2,889	1,018	162	.76	.20	.03	–
Task 18: Facially Expressed Emotion Labeling							
All conditions	1,953	353	56	.78	.09	.02	.59/.60
Happiness	1,409	310	50	.93	.10	.02	–
Sadness	2,221	743	119	.79	.21	.03	–
Anger	1,920	760	121	.75	.22	.04	–
Fear	2,681	903	144	.61	.28	.04	–
Surprise	1,769	549	88	.92	.16	.03	–
Disgust	2,182	673	107	.71	.25	.04	–

Task 16: Facially Expressed Emotion Decision

Facial expressions of happiness and anger are more accurately and easily perceived than those of other emotions (Kessler, Bayerl, Deighton, & Traue, 2002), and so were used as the stimuli here in a typical two-choice RT task. Because idiosyncratic configurational features of a person's face might influence emotion recognition, a neutral expression by the same individual preceded the emotional expression, serving as an anchor for the expression decision.

Procedure. A face with a neutral expression was presented for 500 msec, followed by a blank screen (500 msec) and the same individual's face with either a happy or angry expression. Participants indicated whether the second face showed a happy or angry expression. The face stimuli were taken from the AR Face Database (Martinez & Benavente, 1998), the MACBRAIN Expressive Face Database (www.macbrain.org), and the Karolinska Directed Emotional Faces (KDEF; Lundqvist, Flykt, & Öhman, 1998). This test was considered a speed task.

Results and Discussion. The results for this task and the other two emotion recognition tasks are summarized in Table 6. As expected, and in line with the literature (e.g., Matsumoto et al., 2000), happiness recognition was both faster and more accurate than anger recognition [ts(150) = 15.3 and 6.2, ps < .001, ds = 1.07 and 0.74, respectively]. The correlation of the RTs for anger and happiness recognition was .74; on this basis, there was no evidence for emotion specificity in this task.

Task 17: Emotional Odd-Man-Out

As originally described (Frearson & Eysenck, 1986), the odd-man-out task has eight lights situated in a semicircle as the stimuli. On a given trial, three of the lights are illuminated. Two of these lights are situated physically closer together, and the task is to indicate the light that is farthest away from the other two. This task requires a discrimination between two distances: the distance between the first and second target, and that between the second and third target (Danthiir, Wilhelm, & Roberts, 2007). The present task is based on the same notion, using emotionally expressive faces as the stimuli.

Procedure. In each trial, the stimuli were three faces of different people, positioned in a row. Two of the faces expressed the same emotion, one of which was always the middle face. Hence, the different (odd-man-out) emotional expression was always positioned on the left or the right. The participants had to indicate the face that expressed the odd-man-out expression. Each of the six basic emotions was used as target and as distractor. The face stimuli were taken from the AR Face Database (Martinez & Benavente, 1998), the MACBRAIN Expressive Face Database, the KDEF (Lundqvist et al., 1998), and the Caucasian portraits from the Japanese and Caucasian Facial Expressions of Emotion (JACFEE) and the Japanese and Caucasian Neutral Faces (JACNeuF) databases of Matsumoto and Ekman (1998). This test was intended as a speed task.

Results and Discussion. The results are summarized in Table 6. Performance across the six emotions was very accurate. Trials with faces depicting disgust were comparatively hard, but the mean RTs fluctuated across emotions. Across the six emotions, estimates of unidimensionality and homogeneity were not very high, but they were acceptable for a decent psychometric speed measure. Estimates of homogeneity for the accuracies (α = .46 and ω = .51) indicated a small amount of systematic variance in the data as well.

Task 18: Facially Expressed Emotion Labeling

This task was inspired by an existing test, the Facially Expressed Emotion Labeling task (FEEL; Kessler et al., 2002), which is a modification of the Japanese and Caucasian Brief Affective Recognition Test (JACBART; Matsumoto et al., 2000). FEEL measures individual differences in recognition of the six basic emotions, depicted in both Japanese and Caucasian faces. The task here assessed recognition of the six basic emotional expressions in Caucasian faces and under rapid stimulus presentation.

Procedure. A face with a neutral expression was presented for 1 sec, followed by a mask presented for 500 msec. The mask consisted of three *X*s, centrally displayed. Next, the same face was presented again, displaying one of the six emotional expressions, for 200 msec. After the target face disappeared, a list with six labels for the emotional expressions ("happiness," etc.) was presented. Participants indicated which emotion they had just seen by selecting the appropriate label with the computer mouse. The face stimuli were taken from the AR Face Database (Martinez & Benavente, 1998) and the MACBRAIN Expressive Face Database. This was considered an accuracy task.

Results and Discussion. Apart from surprise and happiness, the accuracies were in a range allowing for sufficient discrimination between people (see Table 6 for details). However, the task was easier than expected, despite the low guessing probability, and the estimates of unidimensionality for accuracy rates were relatively low. Given these suboptimal results, the analysis of RTs seemed a viable alternative. Descriptively, the RTs were similar in length to those for the other speed tasks herein; the estimates of unidimensionality (α = .70, ω = .72) were also somewhat better than for the accuracy data, though still not high. Taking the available evidence together, we conclude that the analysis of RTs for this task is currently better justified than the use of accuracy. Experimental manipulations in future studies might show whether or not this task should be considered a speed or an accuracy task.

GENERAL DISCUSSION

In the present study, we took the first steps toward developing a comprehensive test battery for face cognition. A broad variety of tasks was designed to measure different aspects of individual differences in the processing of unfamiliar and learned faces, and the psychometric quality of all indicators was assessed. The 18 tasks reported here were constructed as measures of face perception, face learning, face recognition, and the recognition of facially expressed emotions. *Face recognition*, as defined in our approach, does not refer to the recognition of preexperimentally familiar or famous faces. Our face recognition tasks were instead designed to measure the recognition of newly learned faces, which we consider to be an essential aspect of face recognition ability. For each element of face cognition, some tasks were intended as speed and others as accuracy measures. A number of tasks were inspired by established experimental effects, such as the part–whole effect. Here, we provide insight into the psychometric properties of these difference measures and their magnitude when obtained from an unusually large sample of participants.

Most tasks fulfilled the criteria for being reliable measures of face cognition. On the whole, they showed acceptable to high estimates of unidimensionality and homogeneity, considering .70 as a general minimum for psychometric acceptability. Only 7 of the 18 tasks did not meet this criterion; of these, no coefficient was below .50, and most were in the upper .50s and .60s. An α of .60 can be considered adequate for a task used for research purposes, especially when it is used together with other indicators in a test battery. Even the tasks with less-than-desirable psychometric properties would be informative and usable in concert with other tasks in a test battery.

In general, our tasks were consistent with the intended—but uninstructed—emphasis on either speed or accuracy; only for Task 18 did the a priori classification have to be revised according to the results. Thus, tasks designed as speed tasks typically showed very high accuracy rates and, on average, considerably shorter RTs than the accuracy tasks. Tasks designed to measure accuracy, on the other hand, yielded between 57% and 90% correct performance, allowing for substantial interindividual variability. Taken together, the 18 tasks presented here proved to be promising indicators for specific aspects of face cognition, especially when combined in a comprehensive test battery. All tasks were designed to take into account the criticisms of some of the available tests and, in addition, to provide a range of measures for assessing various domains of face cognition.

Caucasian faces comprised the stimulus materials in all tasks. In the tasks measuring own-race bias, Chinese faces were also used. Since there are no indications that the underlying processes of face cognition differed across ethnicities, we believe that our findings would generalize to both participants and faces of other ethnicities. It would, of course, be intriguing to attempt a replication of these results with non-Caucasian participants working on own-race stimuli.

Several of the presented tasks (i.e., Tasks 2, 4, 6, 7, 9, and 13) allowed for the isolation of established experimental effects by computing difference measures of two conditions. These difference measures are commonly interpreted as performance costs or benefits from the experimental manipulations, representing specific functions of face cognition. We were able to replicate most of these experimental effects (Tasks 7, 9, and 13, and partly Task 4). Surprisingly, in some tasks we did not find the differences to be of magnitudes comparable to those reported in the literature. In particular, we could not replicate the composite-face effect (Task 2), the distinctiveness effect (Task 4), and the part–whole effect (Task 6). Apart from the specific suggestions given in the discussions of each of these tasks to explain these discrepancies, they may also relate to a more general procedural difference. Here, we used a highly standardized set of stimulus materials that excluded most of the external features of the faces. In contrast, almost all of the previous studies had shown the portrayed person at least with hair, and often with some clothing or background (e.g., Bruce & Valentine, 1985; Tanaka & Sengco, 1997; Valentine, 1991; Young et al., 1987). It certainly merits further investigation why the present study, with a large number of participants and trials, as compared with the typically smaller numbers of samples and trials, did not replicate some of the seemingly well-documented previous findings.

All difference measures—even when they replicated established effects—showed consistently low reliabilities, ranging between .02 and .40. Individual differences in these experimental effects were not stable. Therefore, the usefulness of these difference measures and their theoretical interpretation as indicators of specific face cognition functions, in the context of research concerning individual differences, appears to be limited. The reliability of these difference measures, and hence their usefulness, might be improved by lengthening the tasks. In summary, basing judgments about individual differences in face cognition on indicators of mean differences appears to be inadequate, regardless of the magnitude of the differences.

Clearly, the tasks presented here are not a complete and all-embracing battery for the assessment of face cognition. In particular, these tasks do not measure many of the processes described for faces that are famous or have been familiar for a long time, such as access to biographical information and names, or the processing of such variable face aspects as emotional expression or facial speech. Nevertheless, we hold that the functions under consideration here bear a central and fundamental role for many other aspects of face cognition, and should therefore be regarded first. To our knowledge, the present task compilation is the first multivariate approach for the investigation of perception, learning, and recognition of structural information about faces. We see specific strengths in its multivariate nature. First, precision concerning peoples' abilities on the basis of single tasks is usually rather small. The homogeneity estimates for many of the tasks included here indicate that many tasks, rather than one, would be required for creating highly precise assessments of an individual's ability to process faces, and we cannot see why this should be different in any similar instantiations of tasks. Second, the specificity of single tasks is high and usually unknown. By basing assessments on a broad variety of indicators, task specificity becomes less relevant. The specificity of single tasks also makes it hard to disentangle the effects of familiarity with a task or a task format. Third, and most importantly, one wants to talk about constructs rather than values from a single test when discussing face cognition abilities. The multivariate nature of this task compilation allows for abstracting from the results of individual tasks, making it a first step toward a comprehensive test battery that can be used to investigate face cognition abilities. Establishing the factors related to individual differences in face cognition is a prerequisite for investigating the relationship of face cognition with other cognitive abilities, such as object cognition or general cognitive ability.

AUTHOR NOTE

This research was supported by grants from the State of Berlin (NaFöG) and the Deutsche Forschungsgemeinschaft (Wi 2667/2-2). Portions of this research used the CAS-PEAL face database, collected under the sponsorship of the Chinese National Hi-Tech Program and Isvision Tech. Co.; the Color FERET database of facial images, collected under the FERET program; and the Extended Multimodal Face Database and associated documentation. Correspondence related to this article should be addressed to G. Herzmann, Humboldt University Berlin, Institute for Psychology, Rudower Chaussee 18, D-10099 Berlin, Germany (e-mail: grit.herzmann@cms.hu-berlin.de).

REFERENCES

Aggleton, J. P., Nicol, R. M., Huston, A. E., & Fairbairn, A. F. (1988). The performance of amnesic subjects on tests of experimental amnesia in animals: Delayed matching-to-sample and concurrent learning. *Neuropsychologia*, **26**, 265-272.

Benton, A. L., & Van Allen, M. W. (1968). Impairment in facial recognition in patients with cerebral disease. *Cortex*, **4**, 344-358.

Bonner, L., Burton, A. M., Jenkins, R., McNeill, A., & Bruce, V. (2003). Meet the Simpsons: Top-down effects in face learning. *Perception*, **32**, 1159-1168.

Breen, N., Caine, D., & Coltheart, M. (2000). Models of face recognition and delusional misidentification: A critical review. *Cognitive Neuropsychology*, **17**, 55-71.

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, **73**, 105-116.

Bruce, V., & Valentine, T. (1985). Identity priming in the recognition of familiar faces. *British Journal of Psychology*, **76**, 373-383.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, **77**, 305-327.

Bruyer, R., & Vanberten, M. (1998). Short-term memory for faces: Ageing and the serial position effect. *Perceptual & Motor Skills*, **87**, 323-327.

Burton, A. M., Bruce, V., & Hancock, P. J. B. (1999). From pixels to people: A model of familiar face recognition. *Cognitive Science*, **23**, 1-31.

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, **81**, 361-380.

Busey, T. A. (1998). Physical and psychological representations of faces: Evidence from morphing. *Psychological Science*, **9**, 476-483.

Carbon, C.-C., & Leder, H. (2007). *The microgenesis of early face processing: Features are processed serially, configurations are processed in parallel*. Manuscript submitted for publication.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"—or should we? *Psychological Bulletin*, **74**, 68-80.

Danthiir, V., Wilhelm, O., & Roberts, R. D. (2007). *More evidence for a higher-order model of mental speed: Factor structure and validity of computerized measures*. Manuscript in preparation.

Diamond, A. (1991). Guidelines for the study of brain–behavior relationships during development. In H. S. Levin, H. M. Eisenberg, & A. L. Benton (Eds.), *Frontal lobe function and dysfunction* (pp. 339-378). New York: Oxford University Press.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, **115**, 107-117.

Dingle, K. J., Duchaine, B. C., & Nakayama, K. (2005). A new test for face perception [Abstract]. *Journal of Vision*, **5**(8), 40a.

Duchaine, B., & Nakayama, K. (2005). Dissociations of face and object recognition in developmental prosopagnosia. *Journal of Cognitive Neuroscience*, **17**, 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**, 576-585.

Elliott, R., & Dolan, R. J. (1999). Differential neural responses during performance of matching and nonmatching to sample tasks at two delay intervals. *Journal of Neuroscience*, **19**, 5066-5073.

Ellis, H. D., & Lewis, M. B. (2001). Capgras delusion: A window on face recognition. *Trends in Cognitive Sciences*, **5**, 149-156.

Endl, W., Walla, P., Lindinger, G., Lalouschek, W., Barth, F. G., Deecke, L., & Lang, W. (1998). Early cortical activation indicates preparation for retrieval of memory for faces: An event-related potential study. *Neuroscience Letters*, **240**, 58-60.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review*, **105**, 482-498.

Fast, K., Fujiwara, E., & Markowitsch, H.-J. (2005). *Der Famous Faces Test*. Göttingen: Hogrefe.

Frearson, W., & Eysenck, H. J. (1986). Intelligence, reaction time (RT) and a new "odd-man-out" RT paradigm. *Personality & Individual Differences*, **7**, 807-817.

Freire, A., Lee, K., & Symons, L. A. (2000). The face-inversion effect as a deficit in the encoding of configural information: Direct evidence. *Perception*, **29**, 159-170.

Gaffan, D., Gaffan, E. A., & Harrison, S. (1984). Effects of fornix transection on spontaneous and trained non-matching by monkeys. *Quarterly Journal of Experimental Psychology*, **36B**, 285-303.

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia*, **45**, 32-41.

Hanley, J. R., Young, A. W., & Pearson, N. A. (1991). Impairment of the visuo-spatial sketch pad. *Quarterly Journal of Experimental Psychology*, **43A**, 101-125.

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

Hole, G. J. (1994). Configurational factors in the perception of unfamiliar faces. *Perception*, **23**, 65-74.

Inquisit (2006). Version 2.0.60616 [Computer software]. Seattle, WA: Millisecond Software.

Kessler, H., Bayerl, P., Deighton, R. M., & Traue, H. C. (2002). Facially Expressed Emotion Labeling (FEEL): PC-gestützter Test zur Emotionserkennung. *Verhaltenstherapie & Verhaltensmedizin*, **23**, 297-306.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, **83**, 1198-1202.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska Directed Emotional Faces* [CD-ROM]. Stockholm: Karolinska Institutet, Department of Clinical Neuroscience.

Martinez, A. M., & Benavente, R. (1998). *The AR Face Database* (Tech. Rep. 24). Barcelona: Universitat Autònoma de Barcelona, Computer Vision Center.

Matsumoto, D., & Ekman, P. (1998). Japanese and Caucasian facial expressions of emotions (JACFEE) and neutral faces (JACNeuF).

Matsumoto, D., LeRoux, J., Wilson-Cohn, C., Raroque, J., Kooken, K., Ekman, P., et al. (2000). A new test to measure emotion recognition ability: Matsumoto and Ekman's Japanese and Caucasian Brief Affect Recognition Test (JACBART). *Journal of Nonverbal Behavior*, **24**, 179-209.

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, **6**, 255-260.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Messer, K., Matas, J., Kittler, J., Luettin, J., & Maitre, G. (1999). XM2VTSDB: The Extended M2VTS Database. In *Proceedings of the 2nd International Conference on Audio and Video-Based Biometric Person Authentication (AVBPA '99)*. New York: Springer. Available online at www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/.

Newell, F. N., Chiroro, P., & Valentine, T. (1999). Recognizing unfamiliar faces: The effects of distinctiveness and view. *Quarterly Journal of Experimental Psychology*, **52A**, 509-534.

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **22**, 1090-1104.

Preminger, S., Sagi, D., & Tsodyks, M. (2007). The effects of perceptual history on memory of visual objects. *Vision Research*, **47**, 965-973.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 4. The Advanced Progressive Matrices*. San Antonio, TX: Harcourt Assessment.

Schacht, A., Werheid, K., & Sommer, W. (2008). The appraisal of facial beauty is rapid but not mandatory. *Cognitive, Affective, & Behavioral Neuroscience*, **8**, 132-142.

Searcy, J. H., & Bartlett, J. C. (1996). Inversion and processing of component and spatial–relational information in faces. *Journal of Experimental Psychology: Human Perception & Performance*, **22**, 904-915.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology*, **46A**, 225-245.

Tanaka, J. W., & Sengco, J. A. (1997). Features and their configuration in face recognition. *Memory & Cognition*, **25**, 583-592.

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology*, **79**, 471-491.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology*, **43A**, 161-204.

Valentine, T., & Bruce, V. (1986). The effects of distinctiveness in recognising and classifying faces. *Perception*, **15**, 525-535.

Valentine, T., & Endo, M. (1992). Towards an exemplar model of face processing: The effects of race and distinctiveness. *Quarterly Journal of Experimental Psychology*, **44A**, 671-703.

Warrington, E. K. (1984). *Manual for the Recognition Memory Test*. Windsor: NFER-Nelson.

Wen, G., Bo, C., Shiguang, S., Delong, Z., Xiaohua, Z., & Debin, Z. (2004). *The CAS-PEAL Large-Scale Chinese Face Database and baseline evaluations* (Tech. Rep. JDL-TR-04-FR-001). Beijing: Chinese Academy of Sciences, Joint Research & Development Laboratory for Face Recognition.

Wickham, L. H. V., Morris, P. E., & Fritz, C. O. (2000). Facial distinctiveness: Its measurement, distribution and influence on immediate and delayed recognition. *British Journal of Psychology*, **91**, 99-123.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, **20**, 59-69.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, **81**, 141-145.

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, **16**, 747-759.

Yovel, G., & Duchaine, B. (2006). Specialized face perception mechanisms extract both part and spacing information: Evidence from developmental prosopagnosia. *Journal of Cognitive Neuroscience*, **18**, 580-593.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, **70**, 123-133.

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ω_h. *Applied Psychological Measurement*, **30**, 121-144.

NOTES

[1] Obviously, to investigate own-race bias, we should have added Asian participants as well. However, our intention was not so much to show an own-race bias, but rather to construct two tasks for measuring face perception (Task 4) and learning (Task 9) with Caucasian and Asian faces. Because of the lack of an Asian participant group, we cannot discuss the results with regard to the own-race bias.

[2] The 18 tasks included 36 conditions, resulting in 5,508 possible values for 153 participants. A total of 102 values (1.85%) were set to missing.

[3] The stimuli used to investigate the distinctiveness effect in Tasks 4 and 10 were previously rated by 34 participants (19 female, 15 male; mean age 30.1 years, *SD* 7.5). A total of 330 faces were rated on a 7-point rating scale as to how likely they would be to be recognized in a crowded train station (Wickham, Morris, & Fritz, 2000). On the basis of these ratings, an approximately normally distributed continuum of typical to distinctive faces was selected for Tasks 4 and 10. In both tasks, distinctiveness ratings of faces classified as typical (median 2) were significantly lower than ratings for distinctive faces (median 4) (Mann–Whitney *U* test: *U*s = 78 and 225, respectively; *p*s < .001).

[4] *Nonselected* refers to the fact that the faces were neither all typical nor all distinct, but a random selection of both types.

(Manuscript received April 16, 2007;
revision accepted for publication February 12, 2008.)