Identifizierung überdurchschnittlicher Gesichtserkennungsfähigkeiten:

Modulierende Faktoren und die psychometrische Qualität des CFMT+ sowie des GFMT-S

Dissertation

zur Erlangung des Doktorgrades der Philosophischen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Lara Aylin Petersen

Kiel, September 2021

Erstgutachterin: Prof. Dr. Anja Leue

Zweitgutachter: Prof. Dr. Olaf Köller

Tag der mündlichen Prüfung: 12.01.2022

Durch den Prodekan Prof. Dr. Michael Elmentaler zum Druck genehmigt: 08.02.2022

Ich male die Nasen absichtlich schief, damit die Leute gezwungen sind, sie anzusehen.

Pablo Picasso (1881-1993), spanischer Maler

Inhaltsverzeichnis

Danksagung

An dieser Stelle bedanke ich mich herzlich bei allen, die mich bei der Erstellung dieser Dissertationsschrift fachlich und persönlich unterstützend begleitet haben.

Zuerst danke ich meiner Doktormutter Prof. Dr. Anja Leue für die fachliche Unterstützung, die schnellen Rückmeldungen, die Möglichkeit an ihrem Lehrstuhl zu promovieren und den Raum mich selbst zu verwirklichen. Ebenso danke ich Prof. Dr. Olaf Köller für die Übernahme der Zweitbegutachtung.

Ein besonderer Dank geht an die großartige Martina mit den Gummibären sowie Tina, Dunja und Femke, die mich im „Erfolgsteam“ und im „Glücksbüro“ mit ihrer menschlichen und fachlichen Kompetenz durch den Alltag und den Feierabend begleitet haben.

Ein großes Dankeschön geht auch an meine Arbeitskolleginnen und Ehemalige aus meiner Arbeitseinheit der Psychologie (Franziska, Fee, Saskia und Valentina), der Mathedidaktik am IPN und der PsychProBier&JuDo-Gruppe für die schöne Zeit in ihrer Gemeinschaft: Kuchi-Montag und Kuchen-Sprichwort-Mittwoch werden mich immer an meine Promotionszeit erinnern!

Ich danke den wissenschaftlichen Hilfskräften und Forschungspraktikantinnen, die mich bei den Labor-Datenerhebungen unterstützt haben: Tessa, Selin, Eva, Marie und Sabrina.

Ganz herzlich danke ich all meinen Freunden aus dem Studium und dem Uni-Chor, die mein Leben seit 10 Jahren bereichern: Sophie, Katharina, Christine, Antje, Mii, Jana, Sina, Anna, Ben, Luise, Jakob, Timm, Vivi, Marlene. Ich hoffe noch viele Jahre mit euch schöne Erinnerungen zu schaffen! Zuletzt möchte ich meiner Familie danken – insbesondere meiner Schwester Lea für ihre Unterstützung in allen Lebenslagen und meiner wunderbaren Mutter Regina, die mich so erzogen hat, stets an mich zu glauben und die Großartigkeit in allem zu entdecken.

Zusammenfassung

Die Fähigkeit, Gesichter zu erkennen und Personen korrekt zu identifizieren, erstreckt sich von unterdurchschnittlichen bis überdurchschnittlichen Leistungen. Personen mit überdurchschnittlichen Gesichtserkennungsfähigkeiten werden „*Super-Recognizer*“ genannt und üblicherweise durch ihre Leistungen in Gesichtserkennungstests identifiziert (in der Regel 2 *SD* über dem Mittelwert einer Kontrollstichprobe). Im Fokus dieser Dissertation stehen die Untersuchung von Faktoren, die die individuelle Gesichtserkennungsleistung modulieren, und die Überprüfung der psychometrischen Qualität von zwei Gesichtserkennungstests. Die Untersuchung von Faktoren, die die Gesichtserkennungsleistung modulieren, trägt dazu bei, individuelle Leistungsunterschiede zu verstehen und zu erklären. So wird einerseits das Verständnis über beteiligte kognitive Prozesse in der Forschung erweitert und andererseits können praktische Implikationen abgeleitet werden, unter anderem für den Einsatz von Super-Recognizern in der Polizei. Als Gesichtserkennungstests wurden in dieser Arbeit der „*Cambridge Face Memory Test Long*“ (CFMT+; Russell et al., 2009) und der „*Glasgow Face Matching Test Short*“ (GFMT-S; Burton et al., 2010) gewählt, da diese häufig in der Gesichtserkennungsforschung und auch bei Praxisprojekten der Polizei eingesetzt werden. Trotz der Einsatzhäufigkeit und -dauer finden sich keine bis wenige psychometrische Kennwerte zu diesen diagnostischen Inventaren in der Literatur. Insbesondere bei der Klassifizierung von Leistungsunterschieden müssen Testwerte jedoch eine reliable und valide Aussagekraft haben, um diagnostische Fehlentscheidungen zu vermeiden. In der vorliegenden Arbeit werden die theoretischen Grundlagen der Gesichtserkennung und drei empirische Studien zusammenfassend vorgestellt, deren englischsprachige Publikationsartikel bzw. Manuskripte ebenfalls in der Arbeit zu finden sind.

Im Rahmen von **Studie 1** wurden der Einfluss des Präsentationsmodus, des Alters und des Geschlechts auf die Leistung im CFMT+ und im GFMT-S untersucht. Zudem wurde der

Zusammenhang zwischen der Testleistung im CFMT+ und im GFMT-S analysiert. Es wurden Daten in einer Laborstichprobe (N = 109) und in einer Onlinestichprobe (N = 1435) erhoben. Die Ergebnisse zeigten signifikant bessere Testleistungen im CFMT+ bei Onlinepräsentation, bei Proband:innen zwischen 26 und 35 Jahren sowie bei Frauen. Beim GFMT-S zeigten sich in Post-hoc-Analysen signifikant bessere Testleistungen nur bei der Onlinepräsentation. Proband:innen mit überdurchschnittlichen Testleistungen im CFMT+ erreichten auch signifikant bessere Leistungen im GFMT-S im Vergleich zur Kontrollgruppe. Als erste Einschätzung der Reliabilität wurde Cronbach's Alpha berechnet, welches für den CFMT+ in der Labor- und in der Onlinestichprobe als sehr gut bewertet werden kann. Der GFMT-S zeigte dagegen nur akzeptable Cronbach's Alpha Koeffizienten in beiden Stichproben. Abschließend wurden die Ergebnisse im Rahmen der Konstruktvalidität des CFMT+ und des GFMT-S diskutiert, ebenso in Bezug zur Testfairness.

Im Fokus von **Studie 2** stand die Prüfung der Reliabilität des Online-CFMT+ und des Online-GFMT-S. Im Rahmen einer Test-Retest-Onlinestudie (N = 72) wurden verschiedene psychometrische Koeffizienten in einem integrativen Ansatz beurteilt. Für den CFMT+ wurden sehr gute Werte für die Test-Retest-Reliabilität, Cronbach's Alpha und zwei verschiedene Split-Half-Reliabilitätskoeffizienten mit Spearman-Brown-Korrektur erreicht. Niedrige mittlere Inter-Item-Korrelationen weisen allerdings auf mögliche Effekte heterogenen Stimulusmaterials hin. Darüber hinaus sprechen signifikante Mittelwertsunterschiede zwischen den zwei Messzeitpunkten für Übungseffekte bei einer wiederholten Durchführung des CFMT+ und sind bei der Einschätzung der Konstruktvalidität zu berücksichtigen. Alle Reliabilitätskoeffizienten des GFMT-S erreichten nicht zufriedenstellende Werte, sodass die Testwerte im GFMT-S von Messfehlern beeinflusst sein könnten.

Im Fokus von **Studie 3** standen weitere Faktoren, die die individuelle Gesichtserkennungsleistung modulieren, und die divergente Validität des Online-CFMT+

sowie des Online-GFMT-S. In einer kombinierten Online- und Laborstudie (N = 55) wurde der Zusammenhang der Testleistungen im CFMT+ sowie im GFMT-S zur Aufmerksamkeit, der Verarbeitungsgeschwindigkeit, zum Arbeitsgedächtnis sowie zur selbst eingeschätzten Gesichtserkennungsfähigkeit untersucht. Aufgrund der kleinen Stichprobengröße wird eine Datennacherhebung angestrebt, sodass die Ergebnisse vorläufig und mit Bedacht interpretiert werden müssen. Die Ergebnisse zeigten signifikante mittlere Korrelationen zwischen den Werten der Selbsteinschätzungsfähigkeit und den Leistungen im CFMT+ sowie im GFMT-S. Die kognitiven Maße erreichten nicht signifikante, gering positive (unter anderem verbales Arbeitsgedächtnis) sowie zum Teil negative Korrelationen (unter anderem Verarbeitungsgeschwindigkeit) zur Leistung im CFMT+ und im GFMT-S. Diese Korrelationen können als divergente Validitätsargumente interpretiert werden. Insgesamt zeigten die niedrigen Korrelationen und eine Varianzaufklärung von 35-36% an den Testleistungen im GFMT-S bzw. im CFMT+, dass weitere Einflussfaktoren untersucht werden müssen, um die individuellen Gesichtserkennungsleistungen besser erklären zu können.

Im letzten Kapitel dieser Arbeit werden die Ergebnisse der drei Studien im Forschungskontext diskutiert und kritisch bewertet. Zusammenfassend sprechen die Ergebnisse der drei Studien für eine zufriedenstellende psychometrische Qualität des CFMT+. Der GFMT-S scheint dagegen nicht zur Klassifizierung von Leistungsunterschieden in der Gesichtserkennung geeignet zu sein. Abschließend werden die Limitationen, die Implikationen und der weitere Forschungsbedarf beschrieben.

Abstract

The ability to recognize faces and correctly identify people ranges from far below average to far above average performance. Individuals with far above average face recognition abilities are called “super-recognizers” and are usually classified by their performance in face recognition tests (2 *SD* above the mean of a control sample). This thesis focuses on factors that modulate face recognition performance and on examining the psychometric quality of two face recognition tests. The investigation of factors modulating face recognition performance can help to understand and explain individual performance differences. Thus, the understanding of the involved cognitive processes is broadened in research and practical implications can be derived, e.g., for the employment of super-recognizers in the police. In this thesis, the Cambridge Face Memory Test Long (CFMT +; Russell et al., 2009) and the Glasgow Face Matching Test Short (GFMT-S; Burton et al., 2010) were selected to measure face recognition ability due to their frequent use in face recognition research as well as in police practice projects. However, despite the frequency and duration of use, there are no or only a few psychometric properties for these tests reported in the literature. Especially, when classifying performance differences, the test values must be reliable and valid to prevent diagnostic errors. This thesis presents the theoretical background of face recognition research and three empirical studies, which are summarized below. The publication articles or manuscripts are also included in this thesis.

Study 1 investigated the influence of presentation mode, age, and gender on the test performance in the CFMT+ and in the GFMT-S. In addition, the study examined the relationship of the CFMT+ and the GFMT-S performance. For this purpose, data were collected in a laboratory (N = 109) and in an online sample (N = 1435). The test performance in the CFMT+ was significantly better when presented online, in participants between 26 and 35 years

of age, and in women. For the GFMT-S, post-hoc analyses showed significantly better test performance only for online presentation. Participants with far above average CFMT+ test performance also performed significantly better in the GFMT-S compared to the control group. As a first evaluation of reliability, Cronbach's Alpha was calculated, which can be rated as excellent for the CFMT+ in the laboratory and the online sample. In contrast, the GFMT-S showed only acceptable Cronbach's Alpha coefficients in both samples. Finally, the results were also discussed in the context of construct validity for the CFMT+ and the GFMT-S, including test fairness.

Study 2 focused on testing the reliability of the Online CFMT+ and the Online GFMT-S. The online test-retest study (N = 72) evaluated various psychometric coefficients in an integrative approach. For the CFMT+, the test-retest reliability, Cronbach's Alpha, and two different split-half reliability coefficients with Spearman-Brown correction reached excellent values. However, the low mean inter-item correlations indicate potential effects of heterogeneous stimulus material. In addition, the significant mean differences between the two measurement time points suggest practice effects when the CFMT+ is administered repeatedly. This should be taken into account when evaluating construct validity. All reliability coefficients of the GFMT-S reached unsatisfactory values, indicating measurement errors may influence the test scores in the GFMT-S.

Study 3 focused on factors modulating individual face recognition performance and the divergent validity of the Online CFMT+ and the Online GFMT-S. This combined online and laboratory study (N = 55) investigated attention, processing speed, working memory, and self-report ability of face recognition performance as modulating factors on the test performance in the CFMT+ and the GFMT-S. Due to the small sample size, a follow-up data collection is attempted. Thus, the results are preliminary and must be interpreted with caution. The results showed significant medium correlations between the self-reported face recognition ability and

the CFMT+ and the GFMT-S, respectively. The cognitive measures achieved non-significant, small positive (e.g., verbal working memory), and a few negative correlations (e.g., processing speed) to the CFMT+ and the GFMT-S. These correlations can be interpreted as divergent validity arguments. Overall, the low correlations and a variance elucidation of 35-36% on the test performance in the GFMT-S and the CFMT+, respectively, indicate that more modulating factors should be investigated to explain the face recognition performances.

The final chapter of this thesis discusses and critically evaluates the three studies in the research context. In summary, the results of the studies suggest a satisfactory psychometric quality of the CFMT+, while the GFMT-S does not seem to be suitable for classifying performance differences in face recognition. This thesis closes with a summary of limitations, of implications, and of further research.

1 Einleitung

Gesichter korrekt zu identifizieren ist ein fundamentaler Bestandteil des sozialen Lebens, zum Beispiel in menschlichen Interaktionen (Bruce & Young, 2012). Zudem müssen im Polizeikontext sowohl bekannte Personen als auch unbekannte Personen als Tatverdächtige korrekt identifiziert werden, um eine Strafverfolgung zu ermöglichen. Die Identifizierung von Personen ist jedoch schwierig. Die korrekte Identifizierungsrate von bekannten Personen liegt bei schlechter Bildqualität bei ca. 92% (Bruce, Henderson, Newman, & Burton, 2001). Die korrekte Identifizierungsrate von unbekannten Personen wird deutlich niedriger und heterogener in der Literatur berichtet, unter anderem zwischen 26% (Henderson, Bruce, & Burton, 2001) und 70% (Bruce et al., 2001). Fehlerquoten von bis zu 74% sind im Strafkontext problematisch, da bei einer fehlerhaften Identifizierung eine unschuldige Person verurteilt werden kann (siehe „*Innocent Project*“, https://innocenceproject.org/). Auf der Suche nach Erklärungen für die hohen Fehlerraten bei der Identifizierung unbekannter Personen hat die Gesichtserkennungsforschung in den letzten Jahren verschiedene sogenannte Schätzvariablen als Einflussfaktoren identifiziert (Sporer, Sauerland, & Kocab, 2014). Schätzvariablen sind Faktoren, die sich nur nach einer Wahrnehmungssituation untersuchen lassen, das heißt deren Wirkung erst nachträglich eingeschätzt werden kann. Sporer et al. (2014) unterscheiden Schätzvariablen weiterhin in Zeugen-, Stimulus-, Situations- sowie Beurteilungsvariablen. Beispielsweise wurden Personenmerkmale wie die Ethnie, das Alter oder das Geschlecht als wichtige Zeugen- bzw. Stimulusvariablen identifiziert (Sporer et al., 2014). Die Suche nach Faktoren, die die Gesichtserkennungsleistung modulieren, ist noch nicht abgeschlossen. So ging man bis 2009 davon aus, dass die individuelle Gesichtserkennungsfähigkeit kein bedeutender Einflussfaktor ist und nur Personen mit angeborener oder erworbener Prosopagnosie (Gesichtsblindheit; Duchaine & Nakayama, 2006) mit ihren unterdurchschnittlichen Fähigkeiten eine abweichende Gruppe von der Norm bilden. Mit der

Studie von Russell, Duchaine und Nakayama änderte sich 2009 diese Sichtweise. Russell et al. (2009) stellten fest, dass es auch Menschen mit überdurchschnittlichen Gesichtserkennungsfähigkeiten gibt, die Testergebnisse zwei Standardabweichungen über dem Mittelwert einer Kontrollgruppe erreichen. Russell et al. (2009) nannten Menschen mit diesen überdurchschnittlichen Fähigkeiten „*Super-Recognizer*“ und stellten in ihrer Publikation den „*Cambridge Face Memory Test Long*“ (CFMT+) zur Identifizierung dieser außerordentlichen Fähigkeiten vor. Es gibt noch kein deutschsprachiges, etabliertes Fachwort für Super-Recognizer, sodass dieser englische Begriff im folgenden Text weiterhin verwendet wird. Unter Annahme normalverteilter Gesichtserkennungsfähigkeiten könnten 2% der Menschen in einer Population als Super-Recognizer identifiziert werden. Die Existenz von Super-Recognizern konnte auch von anderen Forscher:innen bestätigt werden (unter anderem Bobak, Pampoulov, & Bate, 2016; Davis, Lander, Evans, & Jansari, 2016). Daher werden die Gesichtserkennungsfähigkeiten seit 2009 als ein Spektrum betrachtet und die individuellen Gesichtserkennungsfähigkeiten könnten den Zeugenvariablen nach Sporer et al. (2014) zugeordnet werden.

Unterschiede in der Gesichtserkennungsfähigkeit können sowohl auf der interindividuellen Ebene im Spektrum als auch auf intraindividueller Ebene innerhalb einer Leistungsklasse betrachtet werden. So berichten einige Studien, dass es innerhalb der Gruppe der Super-Recognizer intraindividuelle Leistungsunterschiede in verschiedenen Gesichtserkennungstests gibt (Bate et al., 2018; Belanova, Davis, & Thompson, 2021; Ramon, Bobak, & White, 2019). Die Erforschung von inter- als auch intraindividuellen Fähigkeiten, Super-Recognizern und Einflussfaktoren ist einerseits wichtig, um das Grundlagenverständnis kognitiver Prozesse bei der Gesichtserkennung zu erweitern, andererseits bringt dieser Forschungsbereich auch praktische Implikationen mit sich. So werden Super-Recognizer bereits unter Polizeibeamt:innen durch Testverfahren identifiziert und in der Praxis eingesetzt.

Polizeibeamt:innen werden zum Beispiel in London (Davis, Forrest, Treml, & Jansari, 2018; Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016) und in Pilotprojekten in Deutschland (zum Beispiel München: Süddeutsche Zeitung, 2019) in Live- oder Video-Identifizierungssituationen eingesetzt. Damit Super-Recognizer einheitlich als solche klassifiziert werden können, muss jedoch eine Definition des Konstrukts festgelegt werden. Während eine Definition des Konstrukts „Super-Recognizer“ noch diskutiert wird (Ramon, 2021; Ramon et al., 2019), gelten bereits das Personengedächtnis und die Personenvergleichsfähigkeiten als wichtige kognitive Prozesse, beziehungsweise Facetten der Gesichtserkennungsfähigkeit (Bate et al., 2018; Bate, Portch, & Mestry, 2021; Ramon, 2021; Ramon et al., 2019). Aufbauend auf der Definition eines Konstrukts benötigt man objektive, reliable und valide diagnostische Inventare, die zur Messung des Konstrukts geeignet sind und auch zur Klassifizierung von Fähigkeitsunterschieden eingesetzt werden können. Bei vielen etablierten Gesichtserkennungstests ist die psychometrische Qualität allerdings noch nicht oder nicht ausreichend geprüft, um ihre Eignung zur Messung von Fähigkeitsunterschieden zu bestätigen. So gibt es kaum psychometrische Befunde zum CFMT+ (Russell et al., 2009), welcher zur Messung des Personengedächtnisses als Hauptinstrument zur Eingangsdiagnostik von überdurchschnittlichen Gesichtserkennungsfähigkeiten eingesetzt wird. Der „*Glasgow Face Matching Test Short*“ (GFMT-S; Burton et al., 2010) wird zur Messung der Personenvergleichsfähigkeiten ebenso in der Super-Recognizer-Forschung eingesetzt (siehe Tabelle 1 in Ramon et al., 2019). Trotz der Einsatzhäufigkeit und -dauer des CFMT+ und des GFMT-S fehlt es an Befunden zur psychometrischen Qualität.

Im Fokus dieser Arbeit standen daher Faktoren, die die Gesichtserkennungsleistung modulieren, sowie die psychometrische Qualität des CFMT+ und des GFMT-S. Das vorliegende Kapitel 1 befasst sich mit den kognitiven Grundlagen der Gesichtserkennung, den

bekannten Einflussfaktoren, den Fähigkeiten der Super-Recognizer sowie den diagnostischen Inventaren CFMT+ und GFMT-S.

1.1 Der kognitive Prozess der Gesichtserkennung

Die kognitive Verarbeitung von Gesichtern (*face processing*) ist in vielen Situationen notwendig. So ist die Informationsverarbeitung grundlegend für die Gesichtserkennung und Personenidentifizierung (*face identity;* zum Beispiel McCaffery, Robertson, Young, & Burton, 2018). Der kognitive Prozess der Gesichtsverarbeitung und -erkennung wird dabei als eigenständiges kognitives System angesehen (Bruce & Young, 1986; Gobbini & Haxby, 2007; Verhallen et al., 2017; Wilmer, Germine, & Nakayama, 2014). Die Gesichtsverarbeitungsfähigkeit wird mit verschiedenen distinkten, aber untereinander korrelierten Aufgaben gemessen (Bruce & Young, 2012; Fysh, 2018), die jeweils verschiedene Facetten der Gesichtsverarbeitung ansprechen. Es können unter anderem die Erkennung eines Gesichts als Gesicht (*face detection*), die Wahrnehmung eines Gesichts (*face perception*), die Erinnerung an ein Gesicht (*face memory*) und der Vergleich von zwei Gesichtern (*face matching*) als zentrale Aufgaben angesehen werden (Bruce & Young, 2012; Fysh, 2018). Diese Aufgabenarten werden in der Literatur auch als Subprozesse von Gesichtskognition (*face cognition*) und unter dem Begriff Gesichtserkennung (*face recognition*) untersucht (Bate & Bennetts, 2014; Fysh, Stacchi, & Ramon, 2020; Wilhelm et al., 2010; Wilhelm, Herzmann, Kunina, & Sommer, 2007). Außerdem wird in der Forschung angenommen, dass ein gemeinsamer Faktor „*f*“ der Gesichtsverarbeitungsfähigkeit in verschiedenen Aufgaben zugrunde liegt (Fysh, 2018; McCaffery et al., 2018; Verhallen et al., 2017). Ergebnisse aus Zwillingsstudien weisen zudem darauf hin, dass die Ausprägung der Gesichtserkennungsfähigkeit genetisch bedingt ist (Shakeshaft & Plomin, 2015; Wilmer et al., 2010).

Der Gesichtsverarbeitungsprozess startet mit der Wahrnehmung eines Gesichts (Bruce & Young, 1986). Dabei werden Gesichter holistisch (ganzheitlich) wahrgenommen (Meinhardt-Injac, Persike, & Meinhardt, 2014; Richler, Floyd, & Gauthier, 2015). Zeitgleich müssen viele visuelle Informationen kognitiv verarbeitet werden (Bruce & Young, 1986; Calder & Young, 2005; Gobbini & Haxby, 2007; Young & Bruce, 2011). Bei der Informationsverarbeitung ist zu berücksichtigen, dass Gesichter eine wichtige soziale Komponente haben (Bruce & Young, 2012; Schweinberger & Schneider, 2014) und dass das Wahrnehmungssystem Gesichter in einer kategorisierbaren Bedeutung erkennen möchte. So lässt sich das kognitive Verarbeitungssystem leicht täuschen, was bekannte Wahrnehmungsphänomene verdeutlichen. Beispielsweise zeigt die „*Thatcher Illusion*“ (zum Beispiel S. 185 in Bruce & Young, 2012), dass das Verarbeitungssystem bei einem Bild mit einem auf dem Kopf stehenden Gesicht, dessen Augen und Mund aber in der erwartungskonformen Richtung angezeigt werden, die Person ohne die Diskrepanz wahrnimmt. Erst beim Drehen des Bildes wird die Diskrepanz zwischen Kopfdrehung und Auge- bzw. Munddrehung vom Verarbeitungssystem bemerkt, sodass die Illusion aufgedeckt wird. Die „*Thatcher Illusion*“ veranschaulicht nicht nur die Funktion des Verarbeitungssystems Gesichter zu erkennen, sondern auch die ganzheitliche Wahrnehmung von Gesichtern.

Die Bedeutung der sozialen Komponente von Gesichtern zeigt sich ebenso im weiteren kognitiven Verarbeitungsprozess. So werden die visuellen Informationen in Kategorien wie zum Beispiel Geschlecht, Ethnie, Alter, emotionale Zustände und Persönlichkeitseigenschaften verarbeitet (Bruce & Young, 1986). Das theoretische Modell von Bruce und Young (1986; siehe auch Young & Bruce, 2011) ist eines der bekanntesten Modelle zur Beschreibung des kognitiven Prozesses im Rahmen der Gesichtserkennung und -identifizierung. Es beschreibt nicht nur, wie Gesichter kategorisiert verarbeitet werden, sondern auch, wie der Prozess unter anderem zwischen bekannten und unbekannten Gesichtern unterscheidet.

Bruce und Young (1986) beschreiben den Prozess dabei als funktionales Modell, das bei bekannten Gesichtern die zu Beginn kategorisierten Informationen in Gesichtseinheiten (*face recognition units*) weiterverarbeitet und anschließend Knotenpunkte (*person identity nodes*) zur Identifizierung aktiviert. Die Knotenpunkte können mit Identitätsinformationen wie zum Beispiel dem Namen einer Person verknüpft sein. Bei der Verarbeitung unbekannter Gesichter werden „*face recognition units*“ nicht aktiviert, sondern die Gesichter zunächst rein visuell und anschließend nur durch das allgemeine kognitive System verarbeitet, in dem unter anderem Prozesse wie die Aufmerksamkeit die Verarbeitung lenken.

Durch die enge Verbindung zu anderen kognitiven Systemen bieten theoretische kognitive Gesichtsverarbeitungsmodelle wie das von Bruce und Young (1986) Ansätze zur Erklärung der inter- und intraindividuellen Unterschiede in der Gesichtserkennungsfähigkeit. Darüber hinaus können mögliche kognitive Einflussfaktoren abgeleitet werden, die im Verarbeitungssystem wirken könnten. So wird beispielsweise das Arbeitsgedächtnis benötigt, um mehrere Informationen parallel zu verarbeiten (Baddeley, 2003; Oberauer, Süß, Wilhelm, & Wittman, 2003), was unter anderem bei der Lösung von Vergleichsaufgaben in Gesichtserkennungstests wichtig ist. Das Kurz- und Langzeitgedächtnis werden benötigt, um Informationen nach der Gesichtsverarbeitung zu speichern und wieder aufzurufen (Baddeley, Hitch, & Allen, 2020), was unter anderem bei der Lösung von Gedächtnisaufgaben in Gesichtserkennungstests relevant ist. Das Modell von Bruce und Young (1986) beschreibt aber nicht nur die Verarbeitung von visuellen Informationen, wie sie zum Beispiel bei Bildern in Gesichtserkennungstests benötigt wird, sondern beschreibt auch die Verarbeitung auditiver Informationen, die über das kognitive System mit den visuellen Informationen verknüpft sind. Die Verknüpfung der visuellen und auditiven Informationsverarbeitung sowie weitere theoretische Annahmen aus Bruce und Young (1986) konnten in den letzten Jahren empirisch gestützt werden. Beispielsweise bestätigte sich die kognitive Unterscheidung von bekannter

und unbekannter Gesichtsverarbeitung (ein Review: Natu & O'Toole, 2011) sowie der Zusammenhang von Gesichts- und Stimmenerkennung (Blank, Anwander, & Kriegstein, 2011; Jenkins et al., 2021; Kramer, Jones, & Gous, 2021). Der Zusammenhang von Gesichts- und Stimmenerkennung wird später in Abschnitt 1.3 weiter thematisiert. Des Weiteren stimmen die Annahmen von Bruce und Young (1986) mit neurologischen Modellen überein, die die aktivierten Gehirnareale bei der Gesichtsverarbeitung von unbekannten und bekannten Personen als getrennte Prozesse beschreiben (Gobbini & Haxby, 2007; Haxby, Hoffman, & Gobbini, 2000). Zusammenfassend kann der kognitive Prozess der Gesichtsverarbeitung und -erkennung als eigenständiges System beschrieben werden, welcher aufgrund seiner Komplexität durch viele Faktoren beeinflusst wird. Einige Einflussfaktoren konnte die Forschung bereits identifizieren, von denen die für diese Arbeit bedeutsamen im nächsten Abschnitt beschrieben werden.

1.2 Einflussfaktoren auf die Gesichtserkennungsleistung

Die menschliche Wahrnehmung, die Verarbeitung von Gesichtern und der Abruf der Informationen aus dem Gedächtnis werden durch verschiedene Faktoren beeinflusst. Auf die Gesichtserkennungsleistung wirken sich vor allem Schätzvariablen wie Situations-, Beurteilungs-, Zeugen- und Stimulusvariablen aus, da diese Variablen in Verbindung mit der Personenidentifizierung stehen (Sporer et al., 2014). Situationsvariablen sind Faktoren, die bei der Wahrnehmung, der Speicherung, dem Behalten und dem Abruf von Informationen wirken, wie zum Beispiel Lichtverhältnisse (Hill & Bruce, 1996) oder die gerichtete Aufmerksamkeit auf andere Objekte in der Wahrnehmungssituation (zum Beispiel „*Weapon Focus Effect*“; Steblay, 1992). Beurteilungsvariablen sind Faktoren, die der Beweiswürdigung dienen, wie zum Beispiel eine verbale Personenbeschreibung (Brown & Lloyd-Jones, 2005). Als Zeugen- und Stimulusvariablen wurden beispielsweise die Passung der Ethnie, des Alters und des Geschlechts von Proband:innen und Zielperson in Gesichtserkennungstests festgestellt (siehe

auch Kapitel 3 in Bruce & Young, 2012). Da die individuelle Gesichtserkennungsleistung seit der bedeutenden Studie von Russell et al. (2009) ebenfalls als Zeugenvariable betrachtet werden kann und die Leistung in Gesichtserkennungstests durch Stimulusvariablen beeinflusst wird (vgl. Sporer et al., 2014), sollen die Zeugen- und Stimulusvariablen im Folgenden weiter thematisiert werden.

Die bisherige Forschung hat gezeigt, dass Personen der eigene Ethnie (*own-race bias* bzw. *cross-race effect;* McKone et al., 2019; Megreya, White, & Burton, 2011; Meissner & Brigham, 2001), der eigenen Altersklasse (*own-age bias;* Herlitz & Lovén, 2013; Susilo, Germine, & Duchaine, 2013) und des eigenen Geschlechts (*own-gender bias;* Megreya, Bindemann, & Havard, 2011; Rhodes & Anastasi, 2012) signifikant besser korrekt identifiziert werden. In einigen neuen Studien zeigten Frauen allerdings auch bei männlichen Stimuli in Tests signifikant bessere Leistungen als Männer (Bobak, Pampoulov et al., 2016; Østergaard Knudsen, Winther Rasmussen, & Gerlach, 2021), sodass eine vom Stimulusgeschlecht unabhängige bessere Gesichtserkennungsfähigkeit bei Frauen möglich scheint. Aufgrund der heterogenen Befunde zum Geschlecht sollte diese Variable als modulierender Faktor noch weiter untersucht werden. Weitere modulierende Zeugen- und Stimulusvariablen sind unter anderem die Einschätzung der Attraktivität, die Zuschreibung von Persönlichkeitseigenschaften (Stereotype) und der emotionale Gesichtsausdruck (siehe Kapitel 3 und 4 in Bruce & Young, 2012).

Die individuelle Gesichtserkennungsfähigkeit als Einflussfaktor auf die Gesichtserkennungsleistung kann als transsituationales, konsistentes und stabiles Konstrukt (*trait*) beschrieben werden, da die Fähigkeit genetisch bedingt (Shakeshaft & Plomin, 2015; Wilmer et al., 2010) und anscheinend kaum bzw. gar nicht trainierbar ist (Dolzycka, Herzmann, Sommer, & Wilhelm, 2014; Hillstrom, Sauer, & Hope, 2011; Towler et al., 2019; White, Kemp, Jenkins, Matheson, & Burton, 2014). Daher könnte es sein, dass sich die individuelle

Gesichtserkennungsfähigkeit auch auf Augenzeugensituationen im rechtswissenschaftlichen Kontext auswirkt (Busey & Loftus, 2007). Die Erkenntnis individueller Fähigkeitsunterschiede ermöglicht darüber hinaus die Erforschung der kognitiven Strukturen und der Einflussfaktoren auf der Suche nach Erklärungen für die Leistungsunterschiede. Wie schon zu Beginn der Einleitung beschrieben, sind überdurchschnittliche Gesichtserkennungsfähigkeiten neben der Forschung auch für die Polizeiarbeit interessant. Könnte die Polizei die eigenen Mitarbeiterressourcen gezielt nutzen, um zum Beispiel Super-Recognizer für die Identifizierung von Tatverdächtigen einzusetzen, kann dies vorteilhaft für die Identifizierungsquote sein, wie Berichte aus England und Deutschland zeigen. So berichtete Ring (2016) am Beispiel der *Metropolitan* Polizei London, dass die Super-Recognizer-Einheit 150 Personen pro Woche identifiziere, während die Computer Gesichtserkennungssysteme nur 10 zuverlässige Identifizierungen in 18 Monaten erbracht hätten. Die Fähigkeiten von Super-Recognizer-Polizeibeamt:innen der *Metropolitan* Polizei London wurden auch wissenschaftlich untersucht und bestätigt (Davis et al., 2016; Robertson et al., 2016). Weitere Zahlen aus Deutschland sprechen ebenso für den erfolgreichen Einsatz von Super-Recognizern. Die Zeitschrift Spiegel berichtete, dass rund die Hälfte von 140 Tatverdächtigen der Stuttgarter „Krawallnacht“ 2021 durch die Super-Recognizer-Polizeibeamt:innen erfolgreich identifiziert wurden (Pointner, 2021). Diese Beispiele zeigen, wie wichtig der Mensch bei der Identifizierung von Tatverdächtigen ist und dass die Kompetenzen von Super-Recognizern ein hohes praxisrelevantes Potenzial haben. Weitere Forschungsergebnisse zu den Fähigkeiten der Super-Recognizer werden im nächsten Abschnitt beschrieben.

1.3 Die Fähigkeiten von Super-Recognizern

Durch die Erkenntnis, dass die Gesichtserkennungsfähigkeit überdurchschnittlich sein kann, wurden viele Experimente der Gesichtserkennungsforschung der letzten Jahrzehnte erneut durchgeführt. So wurde die individuelle Leistungsfähigkeit bei der Bildung von

Experimental- und Kontrollgruppen in Studiendesigns in der früheren Gesichtserkennungsforschung nicht berücksichtigt. Durch den Einbezug individueller Leistungen in bekannte Paradigmen der Gesichtserkennungsforschung möchte man herausfinden, ob beispielsweise Effekte der holistischen Wahrnehmung von Gesichtern oder der Ethnie auch bei Super-Recognizern wirken. Dadurch erhofft man sich, Rückschlüsse auf mögliche Ursachen für die individuellen Unterschiede ziehen zu können. Neben der Definitionsfrage, welche Fähigkeiten einen Super-Recognizer charakterisieren (Bate et al., 2021; Ramon, 2021), versucht die Forschung aufzuzeigen, was Super-Recognizer über die Ergebnisse in klassischen Gesichtserkennungstests hinaus leisten. Aufgrund des jungen Forschungsbereichs gibt es jedoch nicht viele Studien, die die Zusammenhänge zu anderen Fähigkeiten und die Ursachen für die individuellen Leistungsunterschiede erforscht haben.

Eine erste Studie zur holistischen Wahrnehmung von Gesichtern weist darauf hin, dass Super-Recognizer Gesichtsmerkmale, insbesondere Nasen, erfolgreicher in Vergleichsaufgaben verarbeiten als Kontrollproband:innen (Belanova et al., 2021). Belanova et al. (2021) nutzten „*Part-Whole-Effect*"-Aufgaben, bei denen Proband:innen einzelne Gesichtsteile vergleichen müssen, zum Beispiel Nasen oder Augen. Eine Fokussierung von Super-Recognizern auf die Nasen-Region zeigten ebenso zwei *Eyetracking*-Studien (Bobak, Parris, Gregory, Bennetts, & Bate, 2017; Dunn et al., 2021). Allerdings berichteten Belanova et al. (2021) auch intraindividuelle Unterschiede in der Gruppe der Super-Recognizer, die mit den Forschungsergebnissen zu heterogenen Leistungen bei anderen holistischen Gesichtswahrnehmungsaufgaben übereinstimmen (*Inversion*- und *Composite-Face*-Aufgaben: Belanova, Davis, & Thompson, 2018; Bobak, Bennetts, Parris, Jansari, & Bate, 2016; Rezlescu, Susilo, Wilmer, & Caramazza, 2017). Insgesamt weisen die Befunde zur holistischen Wahrnehmung von Gesichtern darauf hin, dass verschiedene kognitive Prozesse die

außergewöhnlichen Fähigkeiten bedingen müssen, da es sowohl intra- als auch interindividuelle Unterschiede in den Leistungen bei verschiedenen Aufgaben gibt.

Weitere Studien haben gezeigt, dass erwachsene Personen mit einem sehr guten Personengedächtnis einem geringeren Ethnien-Effekt unterliegen (Correll, Ma, & Davis, 2021), Kinder signifikant besser identifizieren (Belanova et al., 2018) und sehr gute Leistungen in der Stimmenerkennung zeigen (Jenkins et al., 2021; für Personenvergleichsfähigkeiten siehe Kramer et al., 2021). Des Weiteren hat eine Studie gezeigt, dass Super-Recognizer bei eingeschränkter Sichtbarkeit der Gesichtsmerkmale (zum Beispiel durch Masken oder Sonnenbrillen) immer noch signifikant bessere Leistungen zeigen als Proband:innen mit durchschnittlichen Gesichtserkennungsleistungen (Noyes, Davis, Petrov, Gray, & Ritchie, 2021). Diese Erkenntnis ist besonders in Anbetracht der COVID-19-Pandemielage seit 2020 für die Sicherheitsdienste oder die Polizei relevant, da an vielen Orten eine Mund-Nasen-Bedeckung getragen wird (Robert Koch Institut, 2021). Solche Ergebnisse verdeutlichen noch einmal das Potenzial von Super-Recognizern für den Praxiseinsatz (siehe Abschnitt 1.2).

Insgesamt zeigen die ersten Befunde über die Leistungen von Super-Recognizern, dass diese weniger stark von Effekten bekannter Einflussfaktoren (Ethnie, Alter, Vermummung von Gesichtsmerkmalen etc.) beeinträchtigt sein könnten als Personen mit durchschnittlichen Gesichtserkennungsfähigkeiten. Allerdings fehlt es an weiteren Studien, um die ersten Befunde zu bestätigen. Darüber hinaus sollten andere bekannte und neue Einflussfaktoren untersucht werden, um die zusammenwirkenden kognitiven Prozesse aufzudecken. Da die Gesichtserkennungsfähigkeiten der Proband:innen in allen Studien über Testleistungen klassifiziert werden müssen, wird im folgenden Abschnitt die inventarbasierte Diagnostik von diesen Fähigkeiten beschrieben.

1.4 Inventarbasierte Diagnostik von Gesichtserkennungsfähigkeiten

Bislang gibt es noch keine diagnostische Testbatterie, die einheitlich von Forschenden oder in der Praxis verwendet wird, um individuelle Gesichtserkennungsfähigkeiten zu messen. Die Zusammenstellung mehrerer psychologischer Inventare zu einer Testbatterie ist jedoch notwendig, um die Facetten der Gesichtserkennungsfähigkeiten abbilden zu können (siehe Aufgabenarten in Abschnitt 1.1) und die diagnostische Entscheidung abzusichern. In der Forschung wird neben dem Einsatz von Leistungstests auch der Einsatz von mehrdimensionalen Fragebögen (Selbsteinschätzung) zur Messung der Gesichtserkennungs-fähigkeit diskutiert (unter anderem Bobak, Mileva, & Hancock, 2019). Fragebögen sind allerdings aufgrund der subjektiven Leistungsbeschreibung von Seiten der Proband:innen eher als Ergänzung im Sinne einer Mehr-Ebenen-Diagnostik zu einer objektiveren Testbatterie zu sehen (Beauducel & Leue, 2014), die Verhaltensdaten erfasst. Zudem sollten die korrespondierenden bzw. divergierenden Zusammenhänge zwischen der selbst eingeschätzten Gesichtserkennungsfähigkeit und den Testergebnissen aus standardisierten Inventaren weiter untersucht werden. Aufgrund der Fokussierung der aktuellen Super-Recognizer-Forschung auf diagnostische Inventare mit Gedächtnis- und Vergleichsaufgaben zur Identifizierung der überdurchschnittlichen Gesichtserkennungsfähigkeiten (Bate et al., 2018; Ramon, 2021; Ramon et al., 2019) werden im Folgenden einige diagnostische Inventare aus beiden Kategorien vorgestellt. Dabei konzentriert sich die Vorstellung auf den CFMT+ (Gedächtnisaufgaben) und den GFMT-S (Vergleichsaufgaben), welche in den drei empirischen Studien dieser Arbeit zur Messung der Gesichtserkennungsfähigkeiten verwendet wurden. In der vergangenen Forschung wurden die beiden Tests auch häufig zusammen eingesetzt (unter anderem Belanova et al., 2021; Bobak, Dowsett, & Bate, 2016; Noyes et al., 2021; Satchell, Davis, Julle-Danière, Tupper, & Marshman, 2019).

1.4.1 Inventare zum Personengedächtnis

Der CFMT+ (Russell et al., 2009) ist das bekannteste diagnostische Inventar zur Identifizierung von Super-Recognizern und wird in fast allen Studien zur Eingangsdiagnostik verwendet (Ramon et al., 2019). Der Artikel zur Testvorstellung von Russell et al. (2009) wurde bereits 477-mal zitiert (Information abgerufen am 01.09.2021 über Google Scholar). Im Folgenden werden der CFMT+ und einige kürzlich neu entwickelte, alternative diagnostische Inventare zur Messung vorgestellt.

Cambridge Face Memory Test Long (CFMT+; Russell et al., 2009)

Der CFMT+ ist ein computerbasierter Gesichtserkennungstest zur Messung des Kurzeit-Personengedächtnisses und stellt eine Erweiterung der kürzeren Version CFMT (Duchaine & Nakayama, 2006) dar. Der CFMT wurde ursprünglich zur psychologischen Diagnostik von Prosopagnosie (Gesichtsblindheit) entwickelt, während durch die Erweiterung um 30 schwere Items in einem vierten Testteil der CFMT+ auch im oberen Leistungsbereich (Super-Recognizer) zuverlässig messen soll. Nach einer Instruktionsphase sollen im CFMT+ zunächst sechs männliche Gesichter gelernt werden (Zielgesichter). Die Graustufen-Stimuli fokussieren die inneren Gesichtsmerkmale, indem Haare weggeschnitten sind und keines der männlichen Gesichter mit Bart oder Brille präsentiert wird. Die Proband:innen werden in vier Abrufphasen gebeten, eines der sechs Zielgesichter unter drei Bildern zu identifizieren (siehe Abbildung 1). Die Zielgesichter sind dabei gemäß einer *task-present*-Bedingung immer unter den drei Alternativen zu finden. In den vier Testteilen der Abrufphase wird die Erkennbarkeit der Gesichter stetig schwieriger (siehe Abbildung 1). Sind die Stimuli im ersten Teil noch in identischer Optik zur Lernphase, folgen im zweiten Teil Variationen in der Ansicht oder Belichtung der Gesichter. Im dritten Teil wurde über die Stimuli zusätzlich mit einem Bildbearbeitungsprogramm ein visuelles Rauschen gelegt, sodass die Gesichtsmerkmale noch schwieriger zu erkennen sind. Im vierten Teil mit den 30 zusätzlichen Items werden Stimuli

eingesetzt, bei denen das visuelle Rauschen noch einmal intensiver ist. Zudem werden auch Gesichter von Jugendlichen oder verzerrte Gesichter der Zielgesichter und Distraktoren verwendet. Insgesamt werden in der Abrufphase 102 Items bearbeitet. Die Antworten werden dichotom bewertet (0/1). Für eine richtige Antwort gibt es pro Item einen Punkt und für eine falsche Antwort keinen Punkt, sodass im CFMT+ maximal 102 Punkte erreicht werden können.

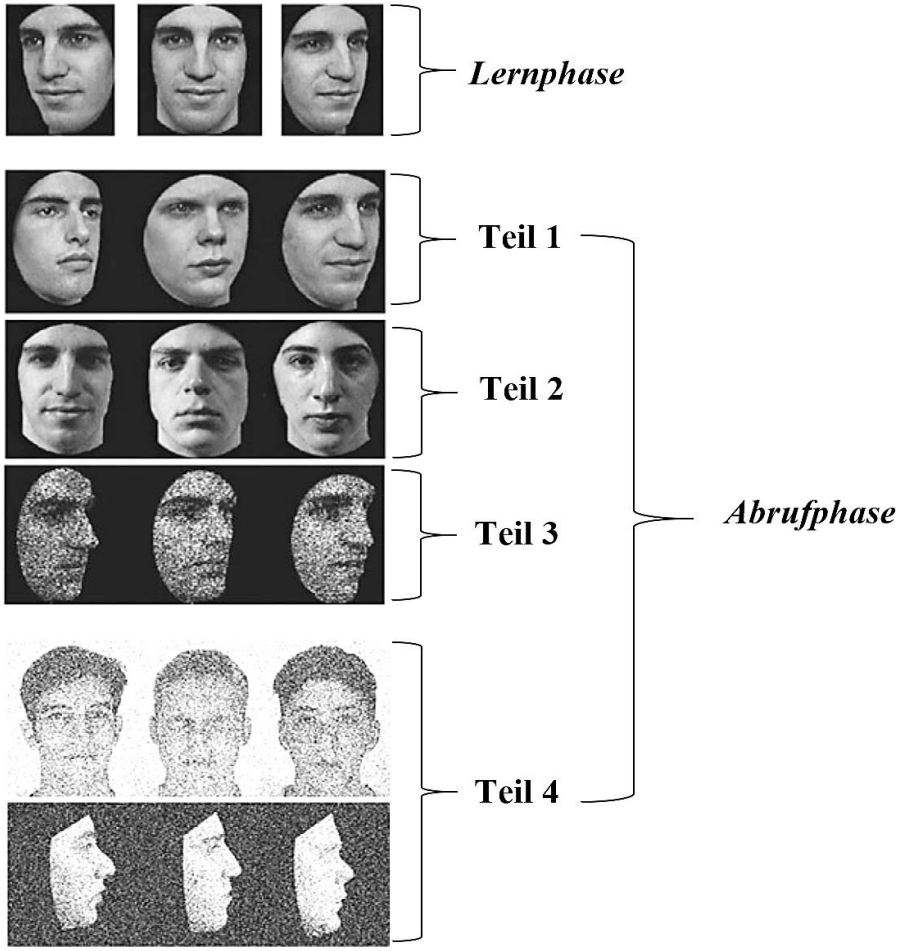


Abbildung 1. Der CFMT+ (Russell et al., 2009) besteht aus einer Lern- und einer Abrufphase. Die Abbildung zeigt den Ablauf der Testphasen mit Beispielitems aus Russell et al. (2009) . Bis zu Teil 3 der Abrufphase ist der Ablauf übereinstimmend mit dem kürzeren CFMT (Duchaine & Nakayama, 2006). Teil 4 kommt nur im CFMT+ vor.

In der Literatur werden sehr schwankende Testmittelwerte für den CFMT+ berichtet, sodass verschiedene Empfehlungen für *Cut-off*-Werte existieren. Testmittelwerte in Stichproben schwanken unter anderem zwischen 70.72 (*SD* = 12.32, Laborstichprobe mit *N* = 254; Bobak, Pampoulov et al., 2016) und 86.62 (*SD* = 9.52, Onlinestichprobe mit *N* = 597; Davis, Bretfelean, Belanova, & Thompson, 2020). Es wäre möglich, dass methodische Faktoren wie der Präsentationsmodus das Testergebnis im CFMT+ beeinflussen. Hierzu fehlt es aber an Untersuchungen. Häufig wird der Empfehlung von Bobak, Pampoulov et al. (2016) gefolgt, die aufgrund ihrer Laborstudienergebnisse von jungen Erwachsenen 95 Rohwerte im CFMT+ als *Cut-off*-Wert für eine überdurchschnittliche Gesichtserkennungsfähigkeit vorschlagen. Jedoch steht eine Überprüfung dieser Empfehlung noch aus, weil bislang keine Normwerte anhand einer alters- und geschlechtsrepräsentativen Normstichprobe für den CFMT+ berechnet wurden. So wäre es für eine einheitliche Interpretation der Testwerte im CFMT+ sinnvoll, eine Normtabelle zu berechnen und standardisierte Normwerte (zum Beispiel T-Werte) zu veröffentlichen. Die Anwendung von Normwerten wird empfohlen, um diagnostische Einzelfallentscheidungen treffen zu können (siehe DIN 33430; Beauducel & Leue, 2014; DIN, 2016).

Weitere Inventare zur Messung des Personengedächtnisses

Der CFMT+ war lange Zeit das einzige diagnostische Inventar, das in allen wissenschaftlichen Studien verwendet wurde, um überdurchschnittliche Fähigkeiten im Personengedächtnis zu messen. So gab es zu Beginn des Dissertationsprojektes 2016 kaum alternative diagnostische Inventare zum CFMT+. Dies änderte sich zuletzt, da seit ca. zwei Jahren vermehrt neue diagnostische Inventare zur Messung des Personengedächtnisses entwickelt werden, um die Testergebnisse im CFMT+ im Sinne einer diagnostischen Testbatterie zu ergänzen oder auch um Alternativen zu bieten. Zum Beispiel wurde 2020 der „*UNSW Face Test*" (Dunn, Summersby, Towler, Davis, & White, 2020) als Alternative zum

CFMT+ vorgestellt, der sowohl in einem ersten Teil das Personengedächtnis sowie in einem zweiten Teil die Personenvergleichsfähigkeit misst. Nur für wenige Gesichtserkennungstests werden so umfassend psychometrische Kennwerte berichtet wie für den „*UNSW Face Test*“ (siehe Dunn et al., 2020) oder den kürzeren CFMT (siehe Wilmer et al., 2012). In der Gesichtserkennungsforschung ist es eher selten, dass Messverfahren psychometrisch umfassend geprüft werden. Für den CFMT+ wurden bislang kaum psychometrische Kennwerte berichtet oder in Bezug auf die Testgüte interpretiert. Das Besondere am „*UNSW Face Test*“ ist nicht nur die Kombination aus Gedächtnis- und Vergleichsaufgaben, sondern auch die Verwendung realitätsnaher Stimuli, zum Beispiel Alltagsfotos von Menschen. Bei klassischen Gesichtserkennungstests wie dem CFMT+ wird vermehrt kritisiert, dass am Computer bearbeitete Stimuli nicht Menschen im realen Leben entsprechen (Bate et al., 2018; Ramon et al., 2019). Allerdings werden Gesichtserkennungstests mit am Computer erzeugten idealen Stimuli benötigt, um die maximale Gesichtserkennungsleistung von Proband:innen ohne Effekte von Bildfarbe (Graustufen-Stimuli vs. farbige Stimuli), Ablenkung durch Haare oder Brille messen zu können. Jedoch kann die externe Validität von diagnostischen Tests wie dem CFMT+ mit idealen Stimuli beeinträchtigt sein (Bate et al., 2018; Ramon et al., 2019). Insbesondere im Zusammenhang mit der Nutzung von Gesichtserkennungstests in Polizeiprojekten sollten auch Items mit hoher externer Validität in einer diagnostischen Testbatterie verwendet werden (Ramon et al., 2019; Thielgen, Schade, & Bosé, 2021). Beispielsweise haben Bate et al. (2018) den „*Models Memory Test*“ (MMT) mit realitätsnahen Stimuli als Ergänzung zum CFMT+ entwickelt. Weitere Messverfahren wurden und werden unter anderem von den Forscher:innen um Prof. Davis von der Greenwich Universität in England entwickelt.

Zusammengefasst hat der CFMT+ durch seine Einsatzhäufigkeit in Studien zum Thema inter- und intraindividuelle Gesichtserkennungsfähigkeiten einen enormen Stellenwert in der

Forschung und auch in der Praxis. Jedoch gibt es kaum Berichte über die psychometrische Qualität. Durch den Einsatz des CFMT+ in dieser Arbeit soll auf Basis von psychometrischen Kennwerten eine erste Einschätzung möglich werden, ob der weitere Einsatz zur Klassifizierung interindividueller Gesichtserkennungsfähigkeiten empfohlen werden kann.

1.4.2 Inventare zum Personenvergleich

Der GFMT-S ist ein häufig eingesetztes Inventar zur Messung der Personenvergleichsfähigkeiten. Der Artikel mit der Testvorstellung von Burton, White und McNeill (2010) wurde bereits 473-mal zitiert (Information abgerufen am 01.09.2021 über Google Scholar). Der GFMT-S wurde sowohl in der allgemeinen Forschung zur Gesichtsverarbeitung als auch mehrmals im Super-Recognizer-Kontext eingesetzt (Belanova et al., 2021; Bobak, Dowsett et al., 2016; Satchell et al., 2019). Im Folgenden wird der GFMT-S beschrieben. Anschließend werden einige kürzlich neu entwickelte, alternative diagnostische Inventare zur Messung des Personengedächtnisses vorgestellt.

Glasgow Face Matching Test Short (GFMT-S; Burton et al., 2010)

Der GFMT-S ist die Kurzversion des GFMT (Burton et al., 2010) und besteht aus den 40 schwierigsten Items der Langversion (GFMT 168 Items). Der GFMT-S kann computerbasiert sowie papierbasiert durchgeführt werden. Nach einer kurzen Instruktion werden den Proband:innen sukzessive 40 Items mit je zwei Stimuli in Graustufen präsentiert. Die Stimuli zeigen sowohl weibliche als auch männliche Gesichter. Anders als beim CFMT+ sind die Gesichter im GFMT-S mit Haaren zu sehen (siehe Abbildung 2). Bei jedem Item soll entschieden werden, ob es sich um Gesichter derselben Person oder von zwei verschiedenen Personen handelt. Dabei sind 20 Stimuli übereinstimmend (*match*) und 20 Stimuli unterschiedlich (*mismatch*). Insgesamt können im GFMT-S maximal 40 Punkte erreicht werden, da für eine richtige Entscheidung ein Punkt vergeben wird. Eine falsche Antwort führt zu 0 Punkten, sodass der GFMT-S dichotom (0/1) bewertet wird. Die Testentwickler Burton et

al. (2010) berichten einen überdurchschnittlichen Testwertebereich ab 37 Rohwertpunkten (N = 194). Diese Empfehlung wurde bislang nicht überprüft. Darüber hinaus gibt es ebenso wie beim CFMT+ keine Normwerttabelle für den GFMT-S mit standardisierten Normwerten zur einheitlichen Interpretation von Rohwerten. Die einheitliche Interpretation von Testwerten und die Verwendung von Normwerten ist wichtig, um diagnostische Entscheidungen im Einzelfall treffen zu können (siehe DIN 33430; Beauducel & Leue, 2014; DIN, 2016).

Abbildung 2. Beispielitems für den GFMT-S aus Burton et al. (2010). (A) Nicht übereinstimmende Personen (*mismatch*). (B) Übereinstimmende Personen (*match).*

Weitere Inventare zur Messung der Personenvergleichsfähigkeiten

Ähnlich zum CFMT+ gab es zu Beginn des Dissertationsprojektes 2016 kaum alternative Personenvergleichstests. Dies hat sich jedoch mit dem Interessenzuwachs am Forschungsgebiet der Super-Recognizer geändert. Ramon (2021) hat die wissenschaftlichen Publikationen zum Thema Super-Recognizer von 2009 bis 2020 gezählt und dabei zeigen können, dass insbesondere seit 2016 die Publikationsgesamtanzahl gewachsen ist (bis 2020 insgesamt 21 Publikationen; Ramon, 2021). Außerdem kann der GFMT-S in manchen Anwendungssituationen für die idealen Stimuli kritisiert werden. Diese zeigen wenig Varianz

im Aussehen der Ziel- und Distraktorenbilder, denn die verwendeten Bilder wurden innerhalb eines Tages aufgenommen (Burton et al., 2010). In der realen Welt würde es mehr Varianz zum Beispiel zwischen einem Passfoto und dem aktuellen Aussehen einer Person geben. Daher wurde als Alternative zum GFMT-S nach derselben Konstruktionsart, aber unter Verwendung realer Passfotos der „*Kent Face Matching Test*“ (KFMT) von Fysh und Bindemann (2018) entwickelt. Ebenfalls mit der Absicht ein anwendungsorientiertes Messverfahren zum Personenvergleich einzusetzen, entwickelten Bate et al. (2018) zwei weitere Inventare. Beispielsweise soll im „*Crowds Matching Test*“ (CMT; Bate et al.,2018) in einem Foto einer Menschenmenge eine Zielperson korrekt identifiziert oder als nicht anwesend vermerkt werden. Vor kurzem wurde nach einer neuen Konstruktionsart zur Messung der Personenvergleichsfähigkeit der „*Oxford Face Matching Test*“ (OFMT) von Stantic et al. (2021) entwickelt. Anders als beim GFMT-S gibt es im OFMT ein Zeitlimit bei der Präsentation der Stimuli (16 Sekunden), sodass die Gesichter nicht beliebig lange auf ihre Übereinstimmung von Proband:innen verglichen werden können. Abschließend kann noch erwähnt werden, dass 2021 eine überarbeitete Version des GFMT, der GFMT 2, mit mehr Variation in den Stimuli in Perspektive, Bildfarbe und Gesichtsausdruck veröffentlicht wurde (White, Guilbert, Varela, Jenkins, & Burton, 2021). Auch für Inventare zur Messung der Personenvergleichsfähigkeit gilt, dass die psychometrische Qualitätsprüfung kein Standard ist und daher zum GFMT-S kaum Befunde in der Literatur zu finden sind.

Zusammenfassend gibt es einige diagnostische Inventare, die die Personenvergleichsfähigkeiten messen, jedoch nur wenige mit überprüfter psychometrischer Qualität. Durch die Überprüfung der psychometrischen Qualität des GFMT-S in dieser Arbeit soll psychometrisch basiert und damit empirisch fundiert eine erste Einschätzung möglich werden, ob der weitere Einsatz zur Klassifizierung interindividueller Gesichtserkennungs-fähigkeiten empfohlen werden kann.

1.4.3 Bewertung der psychometrischen Qualität von Gesichtserkennungstests

Nur wenn die psychometrische Qualität eines diagnostischen Inventars den Standards für psychologische Verfahren entspricht (zum Beispiel American Educational Research Association, 2014), haben die Messergebnisse eine Aussagekraft, um interindividuelle Leistungsunterschiede zuverlässig und valide erfassen zu können. Dies ist besonders wichtig, wenn aufgrund von Testwerten diagnostische Entscheidungen getroffen werden, zum Beispiel die Identifizierung von Super-Recognizern unter Polizeibeamt:innen zur Personalauswahl. Ohne geprüfte reliable und valide Testwerte drohen Fehlentscheidungen (American Educational Research Association et al., 2014). Um überdurchschnittliche Gesichtserkennungsfähigkeiten diagnostizieren zu können, müssen die Leistungen im oberen Punktebereich differenziert werden können. Daher sollten Gesichtserkennungstests die wichtigsten diagnostischen Gütekriterien erfüllen, die beispielsweise in den AERA/APA-Standards (American Educational Research Association et al., 2014), von Beauducel und Leue (2014), von Bühner (2011) oder von Moosbrugger und Kelava (2020) beschrieben wurden. Im Folgenden werden die zentralen Gütekriterien zur Prüfung der psychometrischen Qualität eines diagnostischen Inventars vorgestellt und auf den CFMT+ sowie den GFMT-S bezogen.

Die Objektivität ist ein wesentlicher Bestandteil der Gütekriterien und eine Voraussetzung für die standardisierte Testdurchführbarkeit. Eine Definition in Moosbrugger und Kelava (2020, Seite 18) lautet:

> *„Ein Test ist dann **objektiv**, wenn das ganze Verfahren, bestehend aus Testmaterialien, Testdarbietung, Testauswertung und Interpretationsregeln, so genau festgelegt ist, dass der Test unabhängig von Ort, Zeit, Testleiter und Auswerter durchgeführt werden könnte und für eine bestimmte Testperson bezüglich des untersuchten Merkmals dennoch dasselbe Ergebnis und dieselbe Ergebnisinterpretation liefert.“*

Die Objektivität wird demnach über die Art der Testgestaltung, -durchführung und -auswertung bewertet. Der CFMT+ und der GFMT-S haben konkrete schriftliche Instruktionen für die Proband:innen, werden am Computer durchgeführt und enthalten definierte Auswertungsregeln. Daher können für den CFMT+ und den GFMT-S die Durchführungs- und Auswertungsobjektivität angenommen werden (siehe DIN 33430; DIN, 2016).

Die Reliabilität ist ein weiterer wichtiger Bestandteil der diagnostischen Gütekriterien. In Moosbrugger und Kelava (2020, Seite 27) wird die Reliabilität wie folgt definiert:

> *„Ein Test erfüllt das Gütekriterium der **Reliabilität/Zuverlässigkeit**, wenn er das Merkmal, das er misst, exakt, d.h. ohne Messfehler, misst.“*

Um die Zuverlässigkeit einer Messung durch ein Messverfahren einschätzen zu können, gibt es verschiedene Kennwerte und unterschiedliche Berechnungsmethoden. Beispielsweise können auf Grundlage der klassischen Testtheorie die Koeffizienten Cronbach’s Alpha, Split-Half sowie die Test-Retest-Korrelationen berechnet werden. Des Weiteren können zusätzliche Koeffizienten berechnet werden, die zum Beispiel in die Berechnung von Cronbach’s Alpha einfließen und die wie die mittlere Inter-Item-Korrelation (MIC) ein Maß der Itemhomogenität sind (Bernardi, 1994; Bühner, 2011). Bei Leistungstests wie Gesichtserkennungstests gelten Werte über .90 für Cronbach’s Alpha oder Split-Half-Reliabilität als exzellent, Werte von .80 bis .90 als gut, Werte von .70 bis .80 als akzeptabel und Werte unter .70 als fragwürdig (Bühner, 2011; George & Mallery, 2020). Für den kürzeren CFMT werden einige Reliabilitätskoeffizienten in der Literatur berichtet, zum Beispiel Cronbach’s Alpha von .88 (N = 397, Laborstudie; Verhallen et al., 2017) oder Split-Half-Reliabilität mit Spearman-Brown-Korrektur von .91 (N = 439, Onlinestudie; Wilmer et al., 2012). Für den CFMT+ wurde erst vor kurzem ein Cronbach’s Alpha von .89 (N = 200; Bate et al., 2021) berichtet. Trotz der Einsatzhäufigkeit gibt es keine weiteren Angaben in der Literatur zur Reliabilität des CFMT+. Zum GFMT-S lassen sich dagegen einige wenige Reliabilitätskoeffizienten in der Literatur

finden. So wurde für den GFMT-S ein Cronbach's Alpha von .71 (N = 397; Laborstudie; Verhallen et al., 2017) und eine Test-Retest-Reliabilität von .77 (N = 69; Stantic et al., 2021) berichtet. Stantic et al. (2021) haben ebenfalls festgestellt, dass es zum GFMT-S trotz seiner Einsatzhäufigkeit und -dauer bis heute kaum Befunde zur psychometrischen Qualität gibt.

Das dritte Kriterium für die Bewertung der psychometrischen Qualität diagnostischer Inventare ist die Validität. In Moosbrugger und Kelava (2020, Seite 30) wird die Validität so definiert:

> *„**Validität/Gültigkeit** eines Tests liegt vor, wenn der Test das Merkmal, das er messen soll, auch wirklich misst und nicht irgendein anderes."*

Da es nicht die eine Validität gibt, sollten verschiedene Aspekte untersucht werden, um zu prüfen, ob ein Inventar das intendierte Konstrukt misst. In Moosbrugger und Kelava (2020) wird die Validierung als Prozess der Sammlung von Validitätsargumenten bezeichnet, um „*unterstützende Belege (Evidenz) für die beabsichtigten Testwertinterpretationen zu sammeln und diese einer Prüfung zu unterziehen*" (Seite 535). Sprechen die Validitätsargumente für eine Messung des intendierten Konstrukts, kann ein Testwert auf ein Verhalten in anderen Situationen generalisiert werden (Moosbrugger & Kelava, 2020). Ein wichtiger Aspekt der Validität ist die Konstruktvalidität. Die Konstruktvalidität kann zum Beispiel durch Faktorenanalysen oder konvergente und divergente/diskriminante Korrelationen eingeschätzt werden. Für letzteres werden Korrelationen zwischen Testverfahren berechnet, die ähnliche oder andere Konstrukte messen und ein nomologisches Netzwerk bilden (Campbell & Fiske, 1959; Cronbach & Meehl, 1955). Für die Einschätzung der konvergenten Validität ist es außerdem möglich, ein Konstrukt mit verschiedenen Messmethoden zu untersuchen und die Korrelation zwischen den Messmethoden zu berechnen (zum Beispiel Test vs. Fragebogen). Darüber hinaus ist im Anwendungskontext auch die Kriteriumsvalidität wichtig, um einschätzen zu können, ob ein Testergebnis vergangenes oder zukünftiges Verhalten

vorhersagen kann (Moosbrugger & Kelava, 2020). Zur Validitätseinschätzung des CFMT+ und des GFMT-S gibt es vor allem Berichte über Korrelationen zu anderen Gesichtserkennungstests, die als konvergente Validitätsargumente interpretiert werden können (Korrelationen > .30; Kline, 2005). So ist die Korrelation der Testleistung im CFMT+ beispielsweise am höchsten zu der Leistung in einem anderen Personengedächtnistest, dem „*Models Memory Test*“, mit .65 (*N* = 200; Bate et al., 2018) und .42 zum GFMT-S (*N* = 597; Davis et al., 2020). Außerdem korreliert die Testleistung im CFMT+ positiv mit dem Ergebnis eines Fragebogens zur Selbsteinschätzung der Gesichtserkennungsfähigkeit mit .36 (*N* = 96; SFRS; Bobak et al., 2019). Des Weiteren korreliert die Testleistung im GFMT-S zu anderen Leistungen in Personenvergleichstests wie dem „O*xford Face Memory Test*“ zwischen .46 und .59 (*N* = 69; Stantic et al., 2021).

Neben den Hauptgütekriterien Objektivität, Reliabilität und Validität können Nebengüterkriterien wie zum Beispiel die Testfairness (keine systematische Benachteiligung bestimmter Gruppen), die Testökonomie und die Nützlichkeit wichtige Indikatoren der Testqualität sein (Moosbrugger & Kelava, 2020). Die Durchführung des CFMT+ dauert ca. 25 Minuten und die des GFMT-S ca. 8 Minuten, sodass beide Tests als ökonomisch bezeichnet werden können. Aufgrund des Anwendungskontextes im Rahmen der Polizei ist die Nützlichkeit in Forschung und Praxis gegeben. Die Testfairness des CFMT+ und GFMT-S soll in dieser Arbeit neben der Reliabilität und Validität ebenfalls diskutiert werden.

2 Zielsetzung der Dissertation

In Kapitel 1 wurden die aktuellen Forschungserkenntnisse zum Thema Gesichtserkennungsfähigkeiten vorgestellt. Dabei wurde herausgearbeitet, dass diese Fähigkeiten sowohl inter- als auch intraindividuell unterschiedlich sein können. Die Ursachen für diese Unterschiede sind allerdings noch nicht wissenschaftlich hinreichend geklärt. Aus den bisherigen Studien kann vermutet werden, dass unterschiedliche kognitive Prozesse die Gesichtserkennungsfähigkeit bedingen. Daher gibt es einen hohen Forschungsbedarf zu Faktoren, die die Gesichtserkennungsfähigkeit modulieren. Modulierende Faktoren können dabei beispielsweise aus den Zeugen- und Stimulusvariablen (Sporer et al., 2014) der früheren Gesichtserkennungsforschung hergeleitet werden, zum Beispiel demografische Variablen wie Alter und Geschlecht. Ebenso ist es möglich, dass methodische Variablen wie der Präsentationsmodus eines Tests oder die Messmethode (Konstruktmessung mit einem Leistungstest vs. Fragebogen mit Selbsteinschätzung) die Messung der Gesichtserkennungsfähigkeit beeinflussen. Außerdem können mögliche kognitive Einflussfaktoren auf die individuelle Gesichtserkennungsleistung aus Modellen zu kognitiven Gesichtsverarbeitungsprozessen (unter anderem Bruce & Young, 1986) abgeleitet werden, zum Beispiel kognitive Prozesse wie die Aufmerksamkeit oder das Arbeitsgedächtnis (Logie, Camos, & Cowan, 2020).

Die Gesichtserkennungsfähigkeit wird in der Regel über Leistungstests gemessen, die mit verschiedenen Aufgabenarten die Gesichtserkennung erfassen. Allerdings gibt es noch keine standardisierte, diagnostische Testbatterie zur einheitlichen Klassifizierung von Gesichtserkennungsfähigkeiten. Um Super-Recognizer zu identifizieren, werden in der Forschung aktuell vor allem diagnostische Inventare zum Personengedächtnis und zum Personenvergleich eingesetzt (Bate et al., 2018; Ramon, 2021; Ramon et al., 2019). Sehr häufig werden dabei zur Messung des Personengedächtnisses der CFMT+ (Russell et al., 2009) und

zur Messung der Personenvergleichsfähigkeiten der GFMT-S (Burton et al., 2010) eingesetzt. Diagnostische Inventare sollten die psychologischen Standards zur psychometrischen Qualität erfüllen, um eine Empfehlung zur Messung individueller Gesichtserkennungsfähigkeiten aussprechen zu können. Die wichtigsten Kriterien sind unter anderem die Objektivität, die Reliabilität und die Validität (American Educational Research Association et al., 2014; Beauducel & Leue, 2014; Moosbrugger & Kelava, 2020). Allerdings gibt es so gut wie keine psychometrischen Berichte zur Reliabilität und zur Validität des CFMT+ und nur sehr wenige zum GFMT-S. Darüber hinaus gibt es für den CFMT+ und den GFMT-S keine Normtabellen zur einheitlichen Testwertinterpretation. Aufgrund der Einsatzhäufigkeit und des Forschungsbedarfs an psychometrischen Kennwerten wurden für diese Arbeit der CFMT+ und der GFMT-S zur Messung der Gesichtserkennungsfähigkeit ausgewählt.

Diese Dissertation verfolgte aufgrund der oben dargestellten Forschungslücken drei Ziele: Die Untersuchung verschiedener modulierender Faktoren (demografisch, methodisch und kognitiv) im Kontext interindividueller Gesichtserkennungsfähigkeiten, die Prüfung der psychometrischen Qualität des CFMT+ und die Prüfung der psychometrischen Qualität des GFMT-S. In drei Studien wurden abgeleitet aus der Literatur die Forschungsfragen untersucht. Alle Studien sollten neben der Untersuchung von modulierenden Faktoren auch psychometrische Kennwerte zur Reliabilität des CFMT+ und des GFMT-S berichten. Darüber hinaus sollten die Ergebnisse ebenfalls im Sinne der Konstruktvalidität des CFMT+ und des GFMT-S diskutiert werden. Zusätzlich sollten im Rahmen dieser Arbeit Normtabellen für den CFMT+ und den GFMT-S erstellt werden, damit diese von anderen Anwender:innen zur einheitlichen Testwertinterpretation herangezogen werden können.

In Abbildung 3 sind die Forschungsschwerpunkte der drei Studien im Überblick dargestellt. Wie diese Abbildung zeigt, stehen die drei Studien in einem engen Zusammenhang und bauen aufeinander auf. Alle Studien haben die Reliabilität bzw. Validität thematisiert,

während Studie 1 und 3 noch zusätzlich Forschungsfragen zu Einflussfaktoren (demografisch, methodisch, kognitiv) untersucht haben. In den nachfolgenden Kapiteln werden die drei empirischen Studien mit ihrem theoretischen Hintergrund, der Methodik, den Ergebnissen und einigen Diskussionspunkten zusammenfassend dargestellt. Nach jeder Studienzusammenfassung folgt die englischsprachige Publikation bzw. das englischsprachige Manuskript zur Studie.

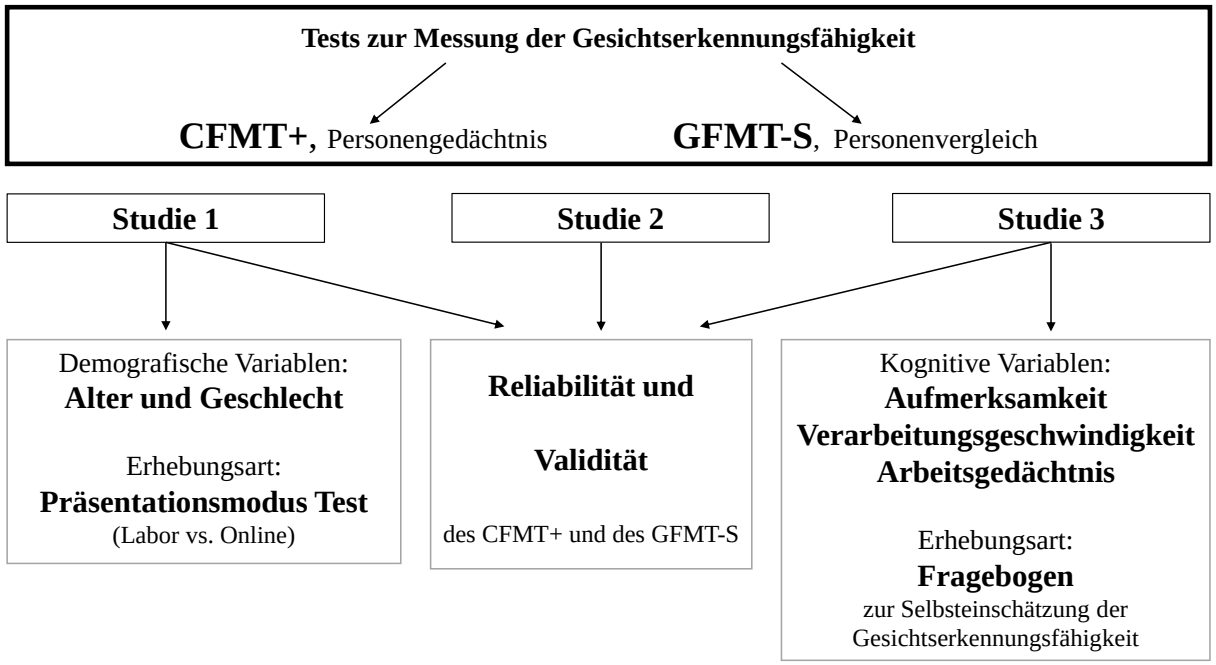


Abbildung 3. Die Forschungsschwerpunkte der Studien 1, 2 und 3 im Überblick.

3 Studie 1 - Extraordinary face recognition performance in laboratory and online testing

Petersen, L. A., & Leue, A. (2021). Extraordinary face recognition performance in laboratory and online testing. *Applied Cognitive Psychology, 35*, 579-589. https://doi.org/10.1002/acp.3805

In diesem Kapitel wird Studie 1 vorgestellt, die unter dem Titel „*Extraordinary face recognition performance in laboratory and online testing*“ in der Fachzeitschrift „*Applied Cogtive Psychology*“ (*Impact Factor* laut Homepage der Fachzeitschrift vom 10.09.2021: 2.01) publiziert wurde. Nachfolgend werden in Abschnitt 3.1 der theoretische Hintergrund, die Methodik und die Ergebnisse zusammenfassend dargestellt sowie einige Diskussionspunkte aus dem Artikel angesprochen. Die veröffentlichte, englischsprachige Volltextversion zur Studie 1 ist in Abschnitt 3.2 zu finden, sodass dort detaillierte Informationen zu den hier zusammengefassten Inhalten nachgelesen werden können.

3.1 Zusammenfassung Studie 1

Theoretischer Hintergrund

Der CFMT+ (Russell et al., 2009) wurde als diagnostisches Inventar zur Einschätzung überdurchschnittlicher Gesichtserkennungsfähigkeiten in Online- und Laborstudien eingesetzt, ohne dass bislang experimentell untersucht wurde, ob sich der Präsentationsmodus auf die Testleistung auswirkt. Analysiert man die Testmittelwerte des CFMT+ in der Literatur, fällt auf, dass sich diese signifikant zwischen Labor- und Onlinestudien unterscheiden. Darüber hinaus gibt es eine große Variation in der Alters- und Geschlechterverteilung der Studienstichproben. Es wäre möglich, dass sich demografische Variablen auf die Gesichtserkennungsleistung im CFMT+ auswirken, da insbesondere das Alter und das Geschlecht von Proband:innen in der vergangenen Forschung schon als Einflussfaktor identifiziert und diskutiert wurden (Herlitz & Lovén, 2013; Rhodes & Anastasi, 2012). So

wurde beispielsweise berichtet, dass Proband:innen im Alter von etwa 30 Jahren im kürzeren CFMT (Susilo et al., 2013) und Frauen im CFMT+ (Bobak, Pampoulov et al., 2016) signifikant bessere Leistungen zeigen als jüngere oder ältere Proband:innen und als Männer. Für den CFMT+ gibt es noch keine Untersuchung des Alters als Einflussfaktor auf die Testleistung. Aufgrund der heterogenen Befunde zum Einfluss des Geschlechts sollte auch das Geschlecht im Zusammenhang mit dem Testergebnis im CFMT+ erneut untersucht werden. Darüber hinaus werden neben dem Personengedächtnis (CFMT+) noch andere Tests zur Messung der Facetten der Gesichtserkennungsfähigkeit eingesetzt, um eine psychometrisch geprüfte, diagnostische Testbatterie zu erarbeiten (Bate et al., 2018; Bobak, Hancock, & Bate, 2016). Mit dem GFMT-S (Burton et al., 2010) wird die Personenvergleichsfähigkeit auch im Super-Recognizer-Kontext untersucht (Bobak, Dowsett et al., 2016; Davis et al., 2016; Satchell et al., 2019). Um eine Testbatterie zusammenzustellen, ist es wichtig, den Zusammenhang von verschiedenen Tests zu untersuchen, Normwerte zur einheitlichen Testwertinterpretation zu verwenden und die psychometrische Qualität der Tests zu prüfen. Für den CFMT+ und den GFMT-S gibt es bislang keine Normtabellen und kaum bis keine Berichte über die psychometrische Qualität. Zusammenfassend wurden in Studie 1 die folgenden vier Forschungsfragen untersucht:

1. Wie wirkt sich der Präsentationsmodus auf die Testleistung im CFMT+ aus?
2. Zeigen Proband:innen im Alter von etwa 30 Jahren signifikant bessere Testleistungen im CFMT+ als jüngere oder ältere Proband:innen?
3. Zeigen Frauen im CFMT+ signifikant bessere Testleistungen als Männer?
4. Zeigen Proband:innen mit überdurchschnittlichen Testleistungen im CFMT+ (Super-Recognizer) signifikant höhere Testleistungen im GFMT-S als eine Kontrollgruppe?

Darüber hinaus sollten psychometrische Kennwerte und Normwerte zum CFMT+ und zum GFMT-S in einer möglichst großen alters- und geschlechtsrepräsentativen Stichprobe berechnet werden.

Methodik

Im Rahmen dieser Studie wurden zwei Stichproben rekrutiert (Erhebungszeitraum: 2016 bis 2018). Die Laborstichprobe (N = 109, 80.73% Frauen, M_{Alter} = 22.39 Jahre, 18 bis 35 Jahre alt) führte die Studie an Computern mit standardisierter Bildgröße in einem Raum des Instituts für Psychologie der Christian-Albrechts-Universität zu Kiel durch. Eine größere zweite Onlinestichprobe (N = 1435, 51.57% Frauen, M_{Alter} = 38.43 Jahre, 18 bis 77 Jahre alt) bearbeitete die Studie in der Regel im Büro oder zu Hause über SoSci-Survey (https://www.soscisurvey.de/) auf individuellen technischen Geräten, deren Displaygröße mindestens 10 Zoll betrug. Alle Proband:innen wurden vor der Studienteilnahme nicht nach ihren Gesichtserkennungsfähigkeiten klassifiziert und bearbeiteten die Tests zum ersten Mal. Der Studienablauf war in beiden Stichproben identisch: Datenschutzaufklärung, Teilnahmezustimmung, Beantwortung demografischer Fragen, Bearbeitung des CFMT+, eine Minute Pause, Bearbeitung des GFMT-S, Beantwortung von Kontrollfragen zu Störungen in der Studiendurchführung und abschließend Anzeige der Testergebnisse. Durch eine Reihe von Kontrollfragen zu den Durchführungsbedingungen wurde die Datenqualität der Online-Proband:innen sichergestellt. Proband:innen, die die Kriterien zur Datenqualität nicht erfüllten (zum Beispiel durch Smartphone-Teilnahme), wurden in der Auswertung nicht berücksichtigt. Um zu überprüfen, ob der Präsentationsmodus einen Effekt auf die Testleistungen hat, wurde aus der Onlinestichprobe eine zur Laborstichprobe alters- und geschlechtsäquivalente Zufallsstichprobe gezogen (ebenfalls N = 109). Die anderen Forschungsfragen wurden anhand der Onlinestichprobe statistisch untersucht sowie wurden erste psychometrische Kennwerte des CFMT+ und des GFMT-S getrennt für die Labor- und Onlinestichprobe berechnet.

Ergebnisse

Sowohl im CFMT+ (Summenwert) als auch im GFMT-S (Summenwert) zeigten die Online-Proband:innen der Zufallsstichprobe (N = 109; CFMT+: M = 82.90, SD = 12.31;

GFMT-S: M = 37.08, SD = 2.98) in t-Tests für unabhängige Stichproben signifikant bessere Testleistungen als die Laborstichprobe (N = 109; CFMT+: M = 74.90, SD = 11.78; GFMT-S: M = 36.09, SD = 3.05). In einer Varianzanalyse mit Alter und Geschlecht als zufällige Zwischensubjektfaktoren und anschließenden Kontrasttests zeigten die 26- bis 35-jährigen Proband:innen signifikant bessere Testleistungen im CFMT+ als jüngere (< 26 Jahre) und ältere Proband:innen (> 35 Jahre). Frauen zeigten signifikant bessere Testleistungen im CFMT+ als Männer. Post-hoc-Analysen für den GFMT-S zeigten keine signifikanten Alters- oder Geschlechtseffekte. Die Pearson-Korrelation zwischen den Testergebnissen im CFMT+ und im GFMT-S betrug in der Laborstichprobe $r(109) = .46$ ($p < .01$) und in der Onlinestichprobe $r(1435) = .49$ ($p < .001$). Dieselbe Pearson-Korrelation zeigte sich zwischen den normalisierten *z*-Werten im CFMT+ und im GFMT-S, die für die Erstellung von T-Normtabellen nach zwei Transformationsmethoden berechnet wurden (siehe *Results* in 3.2 und *Supplement* in 3.3). Proband:innen, die durch überdurchschnittliche Testwerte (normalisierte *z*-Werte) im CFMT+ als Super-Recognizer klassifiziert werden konnten (N = 34 von 1435, 2.37%), zeigten in einem Mann-Whitney-U-Test auch im GFMT-S (normalisierte *z*-Werte) signifikant bessere Testwerte als die anderen Proband:innen (N = 1401). Zusätzlich zu den in der Publikation veröffentlichten Ergebnissen wurde für diesen Dissertationsschrift eine Zufallsstichprobe von N = 34 aus der größeren Kontrollstichprobe (N = 1401) mit SPSS gezogen. Als Ergänzung wurde der normalisierte GFMT-S *z*-Mittelwert der 34 Super-Recognizer (M = 0.88, SD = 0.62) mit dem normalisierten GFMT-S *z*-Mittelwert der Zufallskontrollstichprobe (N = 34; M = 0.03, SD = 0.95) in einem Mann-Whitney-U-Test verglichen. Auch bei gleicher Gruppengröße sind die Super-Recognizer signifikant besser im GFMT-S als die Kontrollstichprobe (U = 875.50, Z = 3.78, $p < .001$).

Der CFMT+ zeigte in der Labor-Administrierung eine sehr gute Reliabilität mit Cronbach's Alpha = .90, *CI* [.87, .93] (vgl. Bühner, 2011). Allerdings zeigte nur die Hälfte aller

Items eine Itemschwierigkeit innerhalb des angestrebten mittleren Bereichs von .20 bis .80 (Bühner, 2011). Der CFMT+ zeigte in der Online-Administrierung wieder eine sehr gute Reliabilität mit Cronbach's Alpha = .92, *CI* [.91, .93], und wieder lagen knapp die Hälfte aller Itemschwierigkeiten im Bereich von .20 bis .80 (vgl. Bühner, 2011). Der GFMT-S erreichte in der Labor-Administrierung eine fragwürdige bis akzeptable Reliabilität mit Cronbach's Alpha = .65, *CI* [.55, .74] (vgl. Bühner, 2011). Fast alle Items des GFMT-S erwiesen sich als sehr leicht mit Itemschwierigkeiten von .71 bis .99 (vgl. Bühner, 2011). In der Onlinestichprobe zeigte der GFMT-S eine akzeptable Reliabilität mit Cronbach's Alpha = .71, *CI* [.69, .73], und sehr leichte Items (Itemschwierigkeiten .80 bis .99).

Diskussion

Insgesamt konnten in dieser Studie signifikante Gruppenunterschiede (Präsentationsmodus, Alter, Geschlecht) für den CFMT+ sowie der Zusammenhang zwischen den Testleistungen im CFMT+ und im GFMT-S gezeigt werden. Damit konnten alle Hypothesen angenommen werden, die aus den vier Forschungsfragen abgeleitet wurden (siehe Theorie). Aufgrund der signifikanten Gruppenunterschiede lässt sich diskutieren, ob gruppenspezifische Normen zur Testwertinterpretation des CFMT+ notwendig sind. Dafür müsste geprüft werden, ob die Gruppenunterschiede auf Konstruktunterschiede zurückzuführen sind oder durch andere Faktoren erklärt werden können, die die Messung beeinflusst haben könnten. Daher sind Gruppenunterschiede auch für die Bewertung der Konstruktvalidität des CFMT+ und des GFMT-S relevant (vgl. Cronbach & Meehl, 1955). Ein Grund für die bessere Testleistung in der Onlinestichprobe könnte die erhöhte Motivation der Proband:innen sein, da diese in der Regel ohne Vergütung und aus persönlichem Interesse an ihren Gesichtserkennungsfähigkeiten an Onlinestudien teilnehmen. Die Psychologiestudent:innen der Laborstichprobe wurden dagegen mit einer „Versuchspersonenstunde" vergütet. Daher sollte der Einfluss der Motivation auf die Gesichtserkennungsleistung weiter untersucht werden

und ggf. Normen auch nach Motivationskontext ergänzend zum Präsentationsmodus angeboten werden. Die Bedeutung von Kontextfaktoren in der psychologischen Diagnostik kann unter anderem in Bennett und Davier (2017) nachgelesen werden. Außerdem können Gruppenunterschiede auf eine Einschränkung der Testfairness hindeuten (Cronbach & Meehl, 1955). Weitere Untersuchungen sind notwendig, um die Frage nach getrennten Normen zu beantworten, unter anderem auch zur Kriteriumsvalidität. Im Rahmen der Konstruktvalidität kann die Korrelation zwischen den Leistungen im CFMT+ und im GFMT-S als konvergentes Validitätsargument interpretiert werden (Campbell & Fiske, 1959). Cronbach's Alpha als erstes Indiz der Reliabilität kann für den CFMT+ als sehr gut und für den GFMT-S nur als fragwürdig bis akzeptabel bewertet werden (vgl. Bühner, 2011). Weitere Forschung zu psychometrischen Kennwerten des CFMT+ und des GFMT-S ist notwendig, um mehr Indizien zur Reliabilität und Validität zu sammeln. Zusammengefasst wurde in Studie 1 deutlich, dass die Messung der Gesichtserkennungsfähigkeit im CFMT+ durch den Präsentationsmodus, das Alter und das Geschlecht moduliert wird. Erstmals wurde eine Diskussion über gruppenspezifische Normen und die psychometrische Qualität des CFMT+ und des GFMT-S angeregt, zu denen erste Ergebnisse berichtet wurden. Mit einer Stichprobengröße von N = 1435 konnte zudem ein robuster Reliabilitätskoeffizient berechnet werden (Charter, 1999). Außerdem können die veröffentlichten T-Normen zukünftig zur einheitlichen Interpretation der Testwerte im CFMT+ und im GFMT-S in der psychometrischen Einzelfalldiagnostik (Beauducel & Leue, 2014) verwendet werden.

3.2 Englischsprachige Publikation der Studie 1

Es folgt die Volltextversion der Publikation zur Studie 1 „*Extraordinary face recognition performance in laboratory and online testing*“, die in der Fachzeitschrift „*Applied Cognitive Psychology*“ erschienen ist.

Petersen, L. A., & Leue, A. (2021). Extraordinary face recognition performance in laboratory and online testing. *Applied Cognitive Psychology*, *35*, 579-589.

https://doi.org/10.1002/acp.3805

Received: 3 July 2020 | Revised: 30 January 2021 | Accepted: 1 February 2021

DOI: 10.1002/acp.3805

RESEARCH ARTICLE

WILEY

Extraordinary face recognition performance in laboratory and online testing

Lara Aylin Petersen | **Anja Leue**

Department of Psychology, Christian-Albrechts-University zu Kiel (CAU), Kiel, Germany

Correspondence
Lara Aylin Petersen and Anja Leue, Psychological Assessment, Differential and Personality Psychology, Department of Psychology, Kiel University, Olshausenstr. 75, 24118 Kiel, Germany.
Email: petersen@psychologie.uni-kiel.de (L.A.P.) and leue@psychologie.uni-kiel.de (A.L.)

Abstract

The Cambridge Face Memory Test Long (CFMT+) is used to investigate extraordinary face recognition abilities (super-recognizers [SR]). Whether lab and online presentation of the CFMT+ lead to different test performance has not yet been investigated. Furthermore, we wanted to investigate psychometric properties of the CFMT+ and the Glasgow face matching test – short (GFMT-S). We report item difficulties, Cronbach's Alpha, and T norms for the CFMT+ and GFMT-S depending on the presentation mode. We analyzed variations of CFMT+ and GFMT-S performance by means of presentation mode, age, and gender. The results showed significant better CFMT+ performance for online presentation, women, and participants between 26 and 35 years of age. For the GFMT-S, significantly higher test performance was shown for online participants exclusively. Overall, we discuss the modulating effects of participants' motivation for online assessments, the necessity of differential norms (e.g., for personnel decisions), and the application of T norms.

KEYWORDS
age, face recognition, gender, norms, online testing, super-recognizers

1 | INTRODUCTION

The seminal study of Russell et al. (2009) revealed individual differences of face recognition ability in the Cambridge Face Memory Test Long Form (CFMT+). Performance differences in face recognition range from very high to very low abilities (Russell et al., 2009). Individuals with extraordinary face recognition performance are called "Super Recognizer" (SR; Bobak, Dowsett, & Bate, 2016; Bobak, Pampoulov, & Bate, 2016; Russell et al., 2009). SRs are particularly associated with extraordinary face memory and face matching abilities (Ramon et al., 2019). The CFMT+ has been originally investigated in laboratory settings (e.g., Russell et al., 2009) and online in recent years (e.g., Davis et al., 2020). However, it has not yet been investigated whether the presentation mode results in different test performance. Especially when tests are used to classify individuals (e.g., police officers; Davis et al., 2018), it is important to use norm values based on tests that fulfill standards in psychological assessment (American Educational Research Association, 2014) and to report psychometric properties of tests.

1.1 | Face memory performance in the CFMT+

Many studies have used the CFMT (i.e., the previous version of the CFMT+; Duchaine & Nakayama, 2006) for face recognition research (DeGutis et al., 2013; McCaffery et al., 2018; Ramon et al., 2019). Good reliability (Bowles et al., 2009; Herzmann et al., 2008) as well as good convergent and divergent validity (Bowles et al., 2009; Dennett et al., 2012; Duchaine & Nakayama, 2006) were reported for the CFMT. Due to ceiling effects of the CFMT (see Russell

TABLE 1 Overview of some studies that reported CFMT+ test results in context of SR research (sorted by presentation mode and author names)

	N, sample description	Presentation mode	Mean CFMT+	SD CFMT+	Mean age[a] (range)	*N* female in %
Belanova et al. (2018)[b]	499, controls (Exp. 1)	online	74.48	6.76	32	72
	315, SRs (Exp. 1)	online	95.50	2.07	33	61
Davis et al. (2020)	597, mixed sample, after screening test[c] (Exp. 1)	online	86.62	9.52	35 (16–74)	60
	415, controls (Exp. 2)	online	75.21	6.40	40	67
	301, SRs (Exp. 2)	online	97.17	1.78	38	74
Satchell et al. (2019)	792, mixed sample	online	83.85	10.47	34	60
Belanova et al. (2018)	28, controls (Exp. 2)	lab	73.10	6.71	41	54
	19, SRs (Exp. 2)	lab	95.20	2.51	39	53
Bobak, Bennetts, et al. (2016)	30, controls	lab	68.40	11.70	32	63
Bobak, Dowsett, and Bate (2016)	20, controls	lab	71.10	10.50	24	50
	20, controls, motivated[d]	lab	71.80	12.70	25	50
Bobak et al. (2018)	96, controls	lab	66.40	13.30	23 (18–35)	74
Bobak, Pampoulov, and Bate (2016)	254, mixed sample	lab	70.72	12.32	21	58
Davis et al. (2018)	7, police experts (SRs)	lab	90.40	8.00	38 (28–48)	29
	92, police identifiers[e]	lab	81.50	10.70	34 (20–52)	23
	62, controls (group 1)	lab	72.20	11.40	27 (18–66)[f]	76[f]
	90, controls (group 2)	lab	69.80	12.50		
Fysh et al. (2020)	146, mixed sample	lab	65.08	11.58	30	54

Note: Mixed sample, no selection of participants' abilities based on CFMT+ performance; controls, subsample without SR participants.
Abbreviations: Exp., experiment; SR, super-recognizer.
[a]The mean age is given in whole years.
[b]Sample information was only available for 282 SRs and 377 controls in Experiment 1.
[c]After a 5-min - *Could you be a Super-Recogniser Test* participants were invited to complete the CFMT+ and GFMT-S.
[d]Besides payment, participants got extra money for every 10% increase in accuracy above 50%.
[e]Members of the London Police SR pool.
[f]Sample characteristics were reported together for control group 1 and 2.

et al., 2009), the CFMT+ has been developed to differentiate performance in the upper ability range and allows for the investigation of extraordinary face recognition performance. Russell et al. (2009) added a subset of 30 more items (also named as trials) to increase the difficulty of the CFMT (Duchaine & Nakayama, 2006). The CFMT+ (Russell et al., 2009) asks participants to memorize six male faces, scaled in gray color without external facial features and to recognize them later in three-face-comparison tasks (see Materials). The use of such laboratory-based tests for the detection of SRs is not without controversy in the research community (Ramon et al., 2019). Although more real-world tests than the CFMT+ are needed (Ramon et al., 2019; Robertson & Bindemann, 2019), such a standard test battery has not yet been established. Initial attempts to develop tests for the detection of SRs with more external validity can be found in the literature (Bate et al., 2018) but have not yet been proofed as standard tests for superior face recognition. Therefore, the investigation of the psychometric properties of the CFMT+ (e.g., reliability, validity) is expected to continue and it is worthwhile to present normative values for the CFMT+ (American Educational Research Association, 2014).

A test score two standard deviations above the mean of a control group is most commonly used in the literature to classify an individual with SR ability (Bobak, Pampoulov, & Bate, 2016; Russell et al., 2009). Some studies reported cut-off values for the CFMT+ based on raw scores to classify people as SRs (e.g., a test score of 95 out of a maximum test score of 102; Bobak, Pampoulov, & Bate, 2016). Mean values of the CFMT+ are heterogeneous across studies and differ significantly between subsamples (see Table 1 for an overview of some studies that reported mean values for the CFMT+ in context of SR research). Analyzing previous studies (see Table 1), the CFMT+ performance differences can be modulated by sample characteristics (e.g., gender and age) and by presentation mode (i.e., online vs. laboratory [lab] study).

Bobak, Pampoulov, and Bate (2016) investigated the face recognition ability of 254 students in a laboratory study and reported a CFMT+ mean of M = 70.72 (SD = 12.32). They suggested a cut-off value of 95 out of 102 points for SR individuals which has been used by other researchers (Davis et al., 2018; Satchell et al., 2019). In an online study, Satchell et al. (2019) reported a CFMT+ score of M = 83.85 (SD = 10.47, N = 792). The CFMT+ mean value in Satchell

et al. (2019)[1] was significantly higher ($t = -16.63$, $df = 1043.5$, $p < .01$, $d = 1.20$; see Table 1) than the CFMT+ mean value in Bobak, Pampoulov, and Bate (2016). In contrast, Germine et al. (2012) did not observe significant CFMT (Duchaine & Nakayama, 2006) differences depending on presentation mode (laboratory vs. online studies). Germine et al. (2012) compared the results of the CFMT in different laboratory versus online studies and for age and gender-matched samples. They found that reliability and variance were similar for data of laboratory and online studies. Furthermore, they did not find any systematic mean differences between laboratory and online CFMT presentations (two of five online samples had significantly higher scores; one laboratory sample had significantly higher scores). To date, no study showed effects of presentation mode (e.g., laboratory vs. online) of the CFMT+. Therefore, the current study aims at analyzing whether the presentation mode modulates the CFMT+ performance.

In prior studies (Table 1), effects of sample characteristics (e.g., age) on CFMT+ performance have not yet been systematically investigated for participants who are younger or older than the people presented in the stimulus material of the CFMT+ (pictures of men, around 30 years of age; Duchaine & Nakayama, 2006). However, the age of the participants and stimulus material can influence the face recognition performance (for a review see Rhodes & Anastasi, 2012). Susilo et al. (2013) reported that participants of about 30 years of age performed better in the CFMT (Duchaine & Nakayama, 2006) than participants of about 20 years of age. In this line, we expected that participants of about 30 years of age perform better than younger or older participants in the CFMT+ (hypothesis 1).

Another sample characteristic that could influence face recognition performance is the gender of the presented faces in the CFMT+ and participants who perform the CFMT+. Gender effects in face recognition research are very heterogeneous (for a review see Herlitz & Lovén, 2013). Some studies show that women generally perform better than men in the CFMT+ (young adults; Bobak, Pampoulov, & Bate, 2016) or CFMT (Susilo et al., 2013). Hence, we expected women to outperform men in the CFMT+ (hypothesis 2).

Discussing modulating effects on CFMT+ performance has also practical implications as the test is used to classify police officers (Davis et al., 2018). In 2015, the Metropolitan Police in London founded a SR special unit to assist the police in screening CCTV (Closed Circuit Television Videos; Davis et al., 2018). As a result, some researchers investigated the abilities of the special unit members to demonstrate their extraordinary face recognition abilities compared to those of a control group (Davis et al., 2016; Robertson et al., 2016). Davis et al. (2018) used the CFMT+ to compare SR (among other police officers) and controls. The police officers significantly outperformed the control groups in the CFMT+ and in other tests.

[1]Participants in Satchell et al. (2019) were derived from the Greenwich database – a large online database started in 2015 by Prof. Davis and his team form the Greenwich University, London.

1.2 | Face matching ability in the GFMT-S

In addition to face memory (measured by the CFMT+), face matching ability is another important component of superior face recognition (Bate et al., 2018; Bobak, Hancock, & Bate, 2016). Further, face matching ability is required in police work, when people in a passport control are compared with their identity documentation pictures (matching task). Therefore, face matching (i.e., comparison of person and pictures) differs conceptually from face memory (i.e., free recall of faces) (Bate et al., 2018; Bobak, Bennetts, et al., 2016; Verhallen et al., 2017). The Glasgow face matching test (GFMT; Burton et al., 2010) has often been used to investigate face matching ability (Davis et al., 2016; Davis et al., 2020; Dowsett & Burton, 2015; Noyes et al., 2018). The short version of the GFMT (GFMT-S; Burton et al., 2010) consists of 40 face pairs. Participants are asked to decide whether the faces of a pair represent the same person or two different persons (see Materials). Although the GFMT-S is frequently used, there are comparatively few reports of the psychometric properties. The GFMT-S has been applied in some SR studies showing that SR mostly perform the GFMT-S with a 95% accuracy or above (Bobak, Dowsett, & Bate, 2016; Davis et al., 2016; Noyes & O'Toole, 2017; Robertson et al., 2016; Satchell et al., 2019). In sum, SRs may not only demonstrate extraordinary face memory abilities, but they may also outperform controls in other cognitive functions such as face matching (Bate et al., 2018; Bobak, Hancock, & Bate, 2016). Therefore, we expected that SRs reach significantly higher scores in the GFMT-S (Burton et al., 2010) than controls (hypothesis 3). Furthermore, we report psychometric properties for the GFMT-S.

1.3 | Summary of research questions and hypotheses

We aimed to investigate effects of the presentation mode (laboratory vs. online) on CFMT+ results. Regarding age and CFMT+ performance, we expected that participants of about 30 years of age perform better than younger or older participants in the CFMT+ (hypothesis 1). Based on prior CFMT+ studies, we expected that women outperform men in the CFMT+ (hypothesis 2). Finally, we expected that SRs reach significantly higher scores in the GFMT-S than controls (hypothesis 3). Furthermore, we wanted to extend the reporting on the psychometric properties of the two frequently used tests in face recognition research.

We adopted and translated the CFMT+ (Russell et al., 2009) and the GFMT-S (Burton et al., 2010) for a German context. In Sample 1, a mixed sample ($N = 109$) performed the CFMT+ in a laboratory setting and we tested the psychometric properties of the adopted tests. In Sample 2, the CFMT+ and the GFMT-S were presented online in a large mixed sample ($N = 1435$) to test the hypotheses. Effects of presentation mode were investigated by comparing the participants of Sample 1 with an age-matched and a gender-matched randomly selected subgroup of Sample 2.

2 | SAMPLE 1 – TEST OF PSYCHOMETRIC PROPERTIES

2.1 | Methods

2.1.1 | Participants

All participants were recruited at Kiel University by advertising flyer and Facebook (Advertising slogan: Have you always wanted to know your ability to recognize people? Have you even noticed that you can easily remember faces?). One hundred and twenty-two individuals participated in this laboratory study. A total of *N* = 109 participants (88 females, 80.73%) with a mean age of 22.39 years (*SD* = 3.70, *range* 18–35 years) were included for data analysis. The participants were excluded when their data indicated very low attention to the test (usage of answer categories 1 = very low or 2 = low attention on a 7-point Likert scale) or when they admitted to be familiar with the CFMT+ or GFMT-S. The participants were not pre-selected for their recognition ability. They mainly reported to be students (91.74%) and all participants were white Caucasian. They had normal or corrected-to-normal vision. All participants received either a monetary reimbursement of 5 € or 1 course credit hour for taking part in this study.

2.1.2 | Materials

The Cambridge Face Memory Test Long Form (*CFMT+*; Russell et al., 2009)

The CFMT+ is an extended version of the computer-based CFMT (Duchaine & Nakayama, 2006) with an additional subset of 30 more difficult trials (see examples of the items in Russell et al., 2009). In a first training section, the test familiarizes participants with six male target faces (gray-scaled) presented from three viewpoints. Subsequently, the participants see three faces (i.e., a triad) and are asked to decide which matches one of the six target faces (correct response = one point; 18 triads in the first section, for examples see Russell et al., 2009). Participants then again review all six target faces for 20 s. In the second section, 30 triads of distractor and target faces differing in position and lightning conditions are presented. After reviewing the six target faces another 20 s, participants watch 24 triads of distractor and target faces with visual noise in the third section. While the CFMT (Duchaine & Nakayama, 2006) ends after this section, the CFMT+ comprises 30 additional triads with visual noise and image variation in age, emotional expression, a fourth viewpoint (profile) and amount of information (cropped and uncropped images of target and distractor faces). Thus, the CFMT+ yields a total score of 102 points (18 + 30 + 24 + 30 = 102 triads with a maximum of one point per triad). The extreme image variation in the fourth section ensures that individuals with extraordinary face recognition ability can be distinguished from those on the upper end of average face recognition ability. This study used an adapted version of the CFMT+. Instructions were given in German language and they were translated close to the original English instruction. Cronbach's Alpha reliability of the CFMT (Duchaine & Nakayama, 2006) has been reported to be .88 (*N* = 397) in a lab study (Verhallen et al., 2017). The CFMT+ is not time-limited and can be typically performed in about 25 min.

The Glasgow Face Matching Test Short Form (*GFMT-S*; Burton et al., 2010)

The short version of the GFMT (Burton et al., 2010) is a face-matching test consisting of 40 pairs of simultaneously presented unfamiliar face images. Half of the face images are given with the same identity and half with different identity. The images show faces in a frontal position, with neutral expression and in grayscale (female and male face pairs, see examples of the items in Burton et al., 2010). Participants are asked whether the faces are of the same person or of two different persons (two answer categories: same or different). Participants obtain one point for the correct answer, which yields a maximum score of 40 points. Cronbach's Alpha reliability of the GFMT-S has been reported to be .71 (*N* = 397) in a lab study (Verhallen et al., 2017). The GFMT-S has no time limit and typically takes about 5–10 min.

2.1.3 | Procedure

All participants performed the online survey (SoSci survey software) of the tests individually on computers or laptops with image size set equal and standardized distance of 40 cm between the table edge and the display. Four persons were tested at the same time. The study was designed according to the ethical principles of the Declaration of Helsinki (World Medical Association, 2013). All participants gave written informed consent at the beginning of the study and then answered questions on demographic data. Subsequently, the CFMT+ and the GFMT-S were administered in a fixed set order with a 1-min pause between both tests. Finally, participants were asked to answer some control questions to assess the quality of the data (e.g., familiarity of the tests, see Participants) before the test results were displayed. Prior to task performance, the first author or research assistant was available for questions in the same room where participants performed the task. We avoided a face-to face sitting position of participants and experimenter to minimize social influences during the testing session.

2.2 | Results

2.2.1 | CFMT+: Distribution and psychometric properties

Normal distribution of the CFMT+ scores was tested with the Kolmogorov–Smirnov test (a-priori alpha level: $p = .20$). The CFMT+ scores differed significantly from normal distribution for the female group, $K\text{-}S(88) = 0.11$, $p = .01$, and for the male group, $K\text{-}S(21) = 0.18$, $p = .09$ (a-priori significance level: $p = .20$ for test of normality, i.e., no deviation from normal distribution). CFMT+ scores for the whole sample were also not normally distributed, $K\text{-}S(109) = 0.10$, $p = .01$. The divergence from the normal distribution is illustrated in the Figure S1 (Supplement, see Supporting Information). Item difficulties ranged from

0.11 to 1.00. Only 52 of the 102 items showed difficulties between the preferred range of .20–.80 (Bühner, 2011). The reliability of the CFMT+ is excellent in the present study (Cronbach's Alpha = .90, CI [.87, .93]; George & Mallery, 2002).

2.2.2 | CFMT+: Mean values and norm

In Table 2 mean values of the CFMT+ for female, male and all participants are presented. For the overall sample, the CFMT+ mean value in Sample 1 (M = 74.90) was significantly higher than the mean value in Bobak, Pampoulov, and Bate (2016) (M = 70.72, t = 3.06, df = 360.5, p < .01, d = 0.34), and significantly lower than the online mean value of Satchell et al. (2019) (M = 83.85, t = −8.24, df = 898.5, p < .01, d = 0.84).

Since the face recognition ability has been considered to be normally distributed (e.g., Bobak, Pampoulov, & Bate, 2016), normalized z-scores were calculated (Rankit normalization method; Soloman & Sawilowsky, 2009; see Figure S9B). We have calculated two T norms based on different methods as different normalization methods result in slightly different T scores. One T norm was calculated using the McCall transformation (McCall, 1939; see Table S2) and the other T norm was calculated using Rankit-based transformation (Soloman & Sawilowsky, 2009; Woerner et al., 2017; see Table S3). We decided to use both methods because the McCall transformation is the original transformation to perform a T norm (McCall, 1939). The Rankit transformation is recommended when calculating comparative values for extreme test scores (Woerner et al., 2017). We provide recommendations for users in the Discussion when to prefer which of both T norms. We preferred T norms because our data were not completely normally distributed. T norms provide the option to correct slight deviations from normal distribution by means of transformation methods like Rankit or McCall transformation (Nunnally & Bernstein, 1994). A lab CFMT+ T score above 70 classifies an individual as SR (i.e., 2 SD above the mean of T scores, which refers to a mean of T scores M = 50, SD = 10). This cut-off value fits to the results of Bobak, Pampoulov, and Bate (2016) because their cut-off value of 95 raw points corresponds to a z-value of 1.97, which corresponds to a T score of 70. In the present sample, two out of N = 109 participants reached a T score above 70 (1.8% of the sample). In sum, the results indicate that the adaption of the CFMT+ to a German sample was successful (e.g., in terms of reliability) and similar to those findings reported in Bobak, Pampoulov, and Bate (2016).

TABLE 2 Sample 1: Laboratory performance of the CFMT+ and GFMT-S raw scores (SD is given in parentheses)

	Female	Male	Total
N	88	21	109
CFMT+ mean (*SD*)	73.77 (11.88)	79.92 (10.32)	74.90 (11.78)
GFMT-S mean (*SD*)	36.07 (3.00)	36.19 (3.36)	36.09 (3.05)

Abbreviations: CFMT+, Cambridge Face Memory Test Long; GFMT-S, Glasgow Face Matching Test Short.

2.2.3 | GFMT-S: Distribution and psychometric properties

Normal distribution of the GFMT-S test scores was tested with the Kolmogorov–Smirnov test (a-priori significance level: p = .20 for test of normality). The GFMT-S scores differed significantly from normal distribution for the female group, K-S(88) = .14, p < .001, and for the male group, K-S(21) = 0.14, p = .20. For the whole sample, the GFMT-S scores were also not normally distributed, K-S(109) = 0.14, p < .001. This deviation from a normal distribution is illustrated in the Figure S4. Item difficulty was low (*range*: .71–.99; Bühner, 2011). Only three out of 40 items showed a medium item difficulty ranging between .20 and .80 (Bühner, 2011). Cronbach's Alpha of the GFMT-S is rather low in the present study (Cronbach's Alpha = .65, CI [.55, .74]; George & Mallery, 2002) although the upper end of the confidence interval for Cronbach's Alpha (.74) is acceptable (George & Mallery, 2002).

2.2.4 | GFMT-S: Mean values, norm, and correlations

Like Burton et al. (2010) or Satchell et al. (2019), we calculated normalized test scores for the GFMT-S. As for the CFMT+, we calculated a T norm for the GFMT-S using the McCall transformation (McCall, 1939; see Table S5) and a norm using the Rankit-based transformation (Soloman & Sawilowsky, 2009; Woerner et al., 2017; see Table S6). Due to a rather low Cronbach's Alpha coefficient, we refrained from suggesting a cut-off value for GFMT-S. The lower the reliability the less accurate the true value of a person is measured with a test (American Educational Research Association, 2014). However, we considered the reliability to be sufficient for an application in Sample 2.

Furthermore, we calculated the Pearson correlation between CFMT+ raw scores and the GFMT-S raw scores (a-priori alpha level: p < .05, two-tailed). The Pearson correlation between the CFMT+ and GFMT-S raw scores is significant ($r(109)$ = .46, p < .01) and represents a medium effect size (Cohen, 1988). Since the raw scores were not normally distributed, we also calculated the Spearman Rank correlation, which represents a large effect size (Spearman's Rho(109) = .52, p < .01). For normalized z-scores (Rankit) of the CFMT+ and the GFMT-S, the Pearson correlation is significant (r (109) = .50, p < .01) and represents a large effect size (Cohen, 1988).

3 | SAMPLE 2 – TEST OF MODULATING EFFECTS

The psychometric properties of the CFMT+ and the GFMT-S in Sample 1 established a sufficient psychometric basis to apply the German adaptations of the CFMT+ and the GFMT-S in Sample 2. In Sample 2, the same tests with the same procedure were presented in an online study. We tested hypotheses 1 to 3 and effects of the presentation mode by comparing samples recruited in Sample 1 (lab version) and 2 (online version).

3.1 | Method

3.1.1 | Participants

Participants were enrolled in several ways (see Participants description of Sample 1 for advertising slogan): through Facebook and newspaper articles (who reported about SR in the police and provided the link to our study as a self-test). Participants were not pre-selected according to their recognition ability. A total of *N* = 3721 participants took part in the online study. A subsample of *N* = 1435 participants (740 females, 51.57%) with a mean age of 38.43 years (*SD* = 11.22, *range* 18–77 years) was included for data analysis. Participants who did not fulfill the criteria for sufficient quality of online data were excluded. The criteria to ensure data quality were operationalized by control questions. Participants were asked to ensure the following criteria, for example, do not participate by mobile phone[2] (see Footnote 2 for *M* and *SD* of CFMT+ and GFMT-S performance from mobile phone participants), do not interrupt answering the items for more than 5 min or report technical problems during task performance. Participants were also excluded when they admitted being familiar with the CFMT+ or GFMT-S. All participants should be white Caucasian (due to the own-race effect; DeGutis et al., 2013). Their self-rated attention to the task should be above 2 (scale ranging from 1 = very low to 7 = very high attention). Similar criteria have already been used by other online studies (see Germine et al., 2012). All participants included in data analysis were white Caucasian and mainly lived in Germany (2.44% lived in Austria and Switzerland). Participants from all 16 German federal states were represented. Compared to the statistical yearbook of 2019, the education level in this sample (*N* = 1435, 79.10% with a university degree or general qualification for university entrance) was twice as high as in Germany (Statistisches Bundesamt, 2019).

3.1.2 | Procedure

The same online survey platform (SoSci survey) as for Sample 1 was used for Sample 2. Thus, Sample 1 and 2 had the same procedure and implementation (see description of Sample 1). Only the context of both samples differed: While participants in Sample 1 took part in a laboratory examination (motivated by course credit points or a small monetary incentive), the online participants in Sample 2 usually filled in the tests at home or at work (e.g., after reading newspaper articles or after internet searching for face recognition studies). The participants were prompted to conduct the study in a quiet environment with reduction of disturbance variables (e.g., turn off mobile phone) and not to use mobile phones for taking part in the study. Furthermore, the online participants were asked to answer control questions at the end of the study to assess the quality of the online data (see sample description for criteria of sufficient online data).

[2]Participation by mobile phone was applied as last data exclusion criterion. Test performance of mobile phone participants (N = 703) in CFMT+: M = 77.91, SD = 12.73; and in GFMT- S: M = 36.19, SD = 3.27.

3.2 | Results

3.2.1 | CFMT+: Distribution and psychometric properties

Normal distribution of the CFMT+ scores were tested with the Kolmogorov–Smirnov test (a-priori alpha level: $p = .20$). The CFMT+ scores differed significantly from normal distribution for the female group, *K-S*(740) = 0.09, $p < .001$, as for the male group, *K-S* (695) = 0.08, $p < .001$. For the whole sample, the Kolmogorov–Smirnov test was also significant, *K-S*(1435) = 0.08, $p < .001$. The divergence from the normal distribution is illustrated in the Figure S7. Item difficulty showed a *range* from .23 to 1.00. A total of 46 of 102 items showed moderate item difficulties between .20–.80 (Bühner, 2011). The reliability of the CFMT+ was excellent in this sample (Cronbach's Alpha = .92, CI [.91, .93]; George & Mallery, 2002).

3.2.2 | CFMT+: Mean values and norm

Table 3 summarizes the descriptives of the CFMT+ scores for female and male participants as well as for the whole sample. Performance of all participants in the online CFMT+ yielded a mean score of 79.96 points (see Table 3). The CFMT+ mean value in Sample 2 was significantly higher compared to the lab mean value (*M* = 70.72) in Bobak, Pampoulov, and Bate (2016) ($t = 11.06$, $df = 1686.5$, $p = < .01$, $d = 0.75$), and significantly lower compared to the online mean value (*M* = 83.85) in Satchell et al. (2019) ($t = -7.54$, $df = 2224.5$, $p = < .01$, $d = 0.33$). The results remain robust when random samples with the same group size as in Bobak, Pampoulov, and Bate (2016) and Satchell et al. (2019) are obtained from Sample 2. As for Sample 1, normalized z-scores were calculated. Based on z-scores, two T norms were calculated – one based on the McCall transformation (McCall, 1939; see Table S8) and another T norm based on the Rankit-Based transformation (Soloman & Sawilowsky, 2009; Woerner et al., 2017; see Table S9). A CFMT+ online T score of 70 classifies an individual as SR (2 SD above the mean of T scores with a mean of T scores = 50, SD of T scores = 10), which is 99 raw points. A total of *N* = 34 of *N* = 1435 participants reached a T score above 70 (2.4% of the sample).

TABLE 3 Sample 2: Online performance of the CFMT+ and GFMT-S raw scores (SD is given in parentheses)

	Female	Male	Total
N	740	695	1435
CFMT+ mean (*SD*)	81.21 (11.78)	78.62 (12.64)	79.96 (12.26)
GFMT-S mean (*SD*)	37.12 (2.96)	36.76 (2.96)	36.94 (2.96)

Abbreviations: CFMT+, Cambridge Face Memory Test Long; GFMT-S, Glasgow Face Matching Test Short.

3.2.3 | GFMT-S: Distribution, psychometric properties, and norms

Normal distribution of the GFMT-S raw scores was tested with the Kolmogorov–Smirnov test. The GFMT-S scores differed significantly from normal distribution for female participants, *K-S*(740) = 0.18, $p < .001$, as for male participants, *K-S*(695) = 0.16, $p < .001$. For the whole sample, the GFMT-S raw scores were also not normally distributed, *K-S*(1435) = 0.17, $p < .001$. This deviation from a normal distribution can be seen in the Figure S10. Item difficulties showed very easy items (*range* .80–.99; Bühner, 2011). The reliability of the online GFMT-S was acceptable (Cronbach's Alpha = .71, CI [.69, .73]; George & Mallery, 2002). Using the same procedure as in Sample 1, we calculated normalized z-scores based on McCall (1939) transformation and Rankit along with two T norms for the GFMT-S (Table S11 for norms using McCall and Table S12 for norms using Rankit). The lower the reliability the less accurate the true value of a person is measured with a test (American Educational Research Association, 2014). Therefore, we refrained from suggesting a cut-off value for GFMT-S for individual diagnostic purposes. However, if applicants want to use a reference scale, they can find norms for the GFMT-S in the Supporting information. The Pearson correlation between CFMT+ raw scores and the GFMT-S raw scores is $r(1435) = .49$, $p = < .001$ (the same for normalized scores) and can be considered as medium (Cohen, 1988). The results were not altered when Spearman Rank correlations were computed (*Rho*(1435) = .49, $p < .01$).

3.2.4 | Presentation mode

To analyze the influence of the presentation mode on the CFMT+ performance, a subsample obtained in Sample 2 ($N = 109$) was randomly selected according to the age range (18–35 years) and gender ratio of Sample 1 (88 women). The mean differences of the CFMT + and the GFMT-S scores between the two samples (different presentation mode) are shown in Table 4. The mean CFMT+ score in Sample 1 (lab: $M = 74.90$) was significantly lower than the mean CFMT+ score in the subsample of Sample 2 (online: $M = 82.90$, $t = -4.90$, $df = 216$, $p < .001$, $d = -0.66$ [Leven's test: $F(1, 216) = 0.23$, $p = .63$]) with *d* rated as a medium effect (Cohen, 1988). Presentation mode effects in GFMT-S were also calculated (post hoc). The mean GFMT-S score was significantly lower in Sample 1 (lab: $M = 36.09$) than the mean GFMT-S score in the subsample of Sample 2 (online: $M = 37.08$, $t = -2.43$, $df = 216$, $p = .02$, $d = -0.33$ [Leven's test: $F(1, 216) = 0.06$, $p = .81$]) with *d* rated as a small effect (Cohen, 1988).

TABLE 4 Means and standard deviations (SD) of the CFMT+ and GFMT-S scores of the laboratory Sample 1 in comparison with the age and gender matched online sample from Sample 2

	Laboratory sample (Sample 1)	Matched online sample (from Sample 2)
N	109	109
CFMT+ mean (*SD*)	74.90 (11.78)	82.90 (12.31)
GFMT-S mean (*SD*)	36.09 (3.05)	37.08 (2.98)

Abbreviations: CFMT+, Cambridge Face Memory Test Long; GFMT-S, Glasgow Face Matching Test Short.

3.2.5 | Age (hypothesis 1) and gender (hypothesis 2)

We used an analysis of variance (ANOVA) with the normalized CFMT + z-scores (Rankit) as a dependent variable. Age and gender were inserted as random between-subject factors to analyze hypotheses 1 and 2 as well as the interaction of age and gender. The age variable was divided into three groups according to hypothesis 1 (see Introduction), so that each participant was assigned to one of the three age groups: ≤ 25 ($N = 174$, female = 100), 26–35 ($N = 477$, female = 230) and ≥ 36 years of age ($N = 784$, female = 410). Levene's test for equality of variances was not significant, $F(5, 1429) = 0.72$, $p = .61$. The ANOVA shows a significant main effect of Age Group, $F(2, 1429) = 48.51$, $p = .02$, partial $\eta^2 = .98$. Follow-up analyses of Age Group differences (contrasts) revealed that adults between the ages of 26–35 performed better than younger adults (≤ 25 years), difference means = −.34, $p < .001$ (Table 5), and better than older adults (≥ 36 years), difference means = −.26, $p < .001$ (Table 5). Younger and older adults did not significantly differ, difference means = .08, $p = .34$ (Table 5). Hypothesis 1 was confirmed because the participants between 26 and 35 years of age outperformed the younger (≤ 25 years) and older (≥ 36 years). A significant main effect of Gender, $F(1, 1429) = 29.16$, $p = .001$, partial $\eta^2 = .78$, with follow-up analyses of gender differences (contrasts) revealed that woman ($M = 0.10$, $SD = .98$) outperformed men ($M = -0.11$, $SD = 1.01$) (difference means = −0.22, $p = .001$). Hypothesis 2 was confirmed. The interaction Age Group and Gender was not significant, $F(2, 1429) = 0.26$, $p = 0.77$, partial $\eta^2 = .00$.

TABLE 5 Mean raw scores and mean normalized z-scores of the CFMT+ and of the raw scores of the GFMT-S for the three age groups (SD is given in parentheses)

Group for analysis	18–25 years	26–35 years	≥36 years
N	174	477	784
CFMT+ mean (*SD*)	78.18 (12.68)	82.05 (11.68)	79.09 (12.37)
Normalized z-CFMT+ mean (*SD*)[a]	−0.14 (1.01)	0.17 (0.98)	−0.08 (0.99)
GFMT mean (*SD*)	36.44 (3.25)	37.33 (2.70)	36.82 (3.03)

Abbreviations: CFMT+, Cambridge Face Memory Test Long; GFMT-S, Glasgow Face Matching Test Short.

[a]Normalized z scores based on Rankit transformation (see results for further description).

Age and gender effects in GFMT-S were calculated post-hoc. We used an analysis of variance (ANOVA) with the normalized GFMT-S z-scores (Rankit) as a dependent variable. Age Group (see above) and Gender were inserted as random between-subject factors. Levene's test for equality of variances was not significant, $F(5, 1429) = 0.90$, $p = .48$. The main effect of Age Group was not significant, $F(2, 1429) = 7.88$, $p = 0.11$, partial $\eta^2 = .89$. The same was true for the Gender main effect of the GFMT-S, $F(1, 1429) = 7.24$, $p = .07$, partial $\eta^2 = .69$. The interaction between Age Group and Gender was also not significant, $F(2, 1429) = 0.98$, $p = .37$, partial $\eta^2 = .00$.

3.2.6 | GFMT-S: Relationship to CFMT+ (hypothesis 3)

To analyze whether the SRs performed significantly better in the GFMT-S than participants of the control group, the sample was divided into two groups based on their normalized CFMT+ z-scores (Rankit; greater or less than 2 *z*). The SR group ($N = 34$) had a GFMT-S z score mean of 0.74 ($SD = 0.62$) and the control group ($N = 1401$) had a GFMT-S z score mean of −0.04 ($SD = 0.94$). The SR group had a CFMT+ raw score mean of 99.71 ($SD = 0.87$) and a GFMT-S raw score mean of 39.06 ($SD = 1.39$). The control group had a CFMT+ raw score mean of 79.48 ($SD = 12.02$) and a GFMT-S raw score mean of 36.90 ($SD = 2.97$). A non-parametric Mann–Whitney-U-Test was conducted because the statistical assumption of equality of variances for an independent t-test was not fulfilled. The group difference of the GFMT-S z mean scores was significant ($U = 11{,}971$, $Z = -5.01$, $p = < .001$), which supports hypothesis 3.

3.3 | Discussion

In two samples we investigated four research questions on the CFMT + (Russell et al., 2009). First, findings show that the presentation mode (lab vs. online) influences the CFMT+ test performance. Online participants reached significantly higher test scores than lab participants. Second, participants between 26 and 35 years of age performed better than younger (≤25 years) or older participants (≥36 years) in the CFMT+. Third, women performed better in the CFMT+ than men. Fourth, participants with above-average CFMT+ test performance also performed significantly better in the GFMT-S (Burton et al., 2010). Furthermore, we present psychometric properties for the CFMT+ and GFMT-S depending on presentation mode. Theoretical and practical implications are discussed below.

3.3.1 | Presentation mode, age, and gender: Are differentiated norms for the CFMT+ necessary?

Germine et al. (2012) did not find any systematic influence of the presentation mode in a data comparison of different studies on CFMT performance (Duchaine & Nakayama, 2006). In contrast to Germine et al. (2012), this paper shows a significantly higher CFMT+ mean for online participants compared to lab participants. The CFMT and CFMT+ measure the same construct (face memory) with the same material. The only difference between the CFMT and the CFMT+ is the higher number of items in the CFMT+ (additional section with more difficult items). In our study, the CFMT+ was proved to be reasonably reliable in the lab and online mode. Germine et al. (2012) considered sample characteristics (e.g., age and gender) as modulating factors on CFMT test performance. In our study, both samples for the presentation mode analyses were comparable in age range and gender ratio. Furthermore, Germine et al. (2012) used online samples from the website testmybrain.org. Therefore, their participants should have been similarly interested or motivated in the study topic as our participants of the online sample (i.e., study participation after internet research or reading newspaper articles). The question remains whether the better CFMT+ online performance can be explained by other variables such as increased participant motivation or the online participants' approach to the survey. Satchell et al. (2019) suggested an influence of motivation or interest in research in their discussion of the high mean values of the CFMT+. It is possible that lab participants (often students; credit points or money for participation) and online participants (participate due to their interests in the topic) are differently motivated to perform tests. If further research would show that motivation or interests of the participants have an impact on CFMT+ performance, the norms might need to be differentiated by motivation context.

Moreover, we recommend classifying SR individuals based on norm values not raw scores because raw scores do not have any references. Moreover, users of the norms should also note that the type of transformation (Rankit vs. McCall) that has been applied to perform T norms might influence comparative values in lab and online samples (Tables S2, S5, S8, and S11 for norms calculated with McCall; Tables S3, S6, S9, and S12 for norms calculated with Rankit). Since different statistical methods for calculating T norms generate slightly different T scores in relation to the associated raw scores (e.g.,: the T score for the raw score 50 in CFMT+ in the S2 and S3; T = 28 and T = 29 for the same raw score), we calculated two T norms for the CFMT+ and GFMT-S (each for online and lab). When using norms in the context of comparison with other tests providing T norms (e.g., personality questionnaires), it is important to compare exclusively those T scores that were calculated with the same statistical method (e.g., when a test uses the McCall transformation for calculating T scores, the CFMT+ and GFMT+ norm in S1–S4 should be applied and not the T scores with Rankit transformation). In addition, it is convenient to use the T norms for the CFMT+ when other tests also offer T norms so that test values can be quickly compared on a common scale.

Furthermore, the significant age group main effect and significant contrasts for the CFMT+ (hypothesis 1) correspond to results of Susilo et al. (2013) for the CFMT (Duchaine & Nakayama, 2006). We also provide evidence of gender differences in online CFMT+ performance. Comparable to Bobak, Pampoulov, and Bate (2016), women reached higher test scores in the CFMT+ than men (hypothesis 2 confirmed). However, the CFMT+ z score mean group differences are

rather small (cf. Table 5). Nevertheless, considering the large effect sizes in our study (after Cohen, 1988; age: partial η^2 = .98, gender: partial η^2 = .78) it seems appropriate to test and develop age-specific and gender-specific norms. Whether CFMT+ performance differences depending on age and gender are related to criterion differences of face recognition remains to be part of future research.

The investigation of group differences (e.g., age and gender) is not only important in the context of research questions and practical implications, but also for examining construct validity (Cronbach & Meehl, 1955) and test fairness. Furthermore, another important aspect of construct validity refers to the comparison of the CFMT+ with other tests in terms of convergent and divergent validity.

3.3.2 | Construct validity: The relationship of the CFMT+ and GFMT-S

The CFMT+ was developed to test face recognition ability and to detect SR more accurately. Prior SR research has not yet investigated a complete set of the cognitive abilities that contribute to the extraordinary face recognition performance of SRs. Besides the superior face recognition ability, the face matching ability could also be above average in SRs (e.g., Bate et al., 2018; Bobak, Dowsett, & Bate, 2016; Robertson et al., 2016). The GFMT-S is a very frequently used test to measure face matching ability. The Pearson correlation of the CFMT+ and GFMT-S (Rankit normalized z-scores, r = .49 and .50) can be considered as evidence of convergent validity. The correlation is not so high that the two tests measure the same construct, but also not so low that they measure completely different constructs (Campbell & Fiske, 1959). Both tests seem to be important for testing face recognition ability and to elucidate the underlying superior cognitive abilities of SRs.

In both samples, the CFMT+ proved to be an instrument of excellent reliability (Cronbach's Alpha = .90 and .92) with acceptable item difficulties (excluding exercise items). These psychometric properties facilitate the CFMT+ as a suitable instrument for the assessment of superior abilities (SR). Contrary to the CFMT+, the test quality of the GFMT-S should be rated worse. In both samples, reliability of the GFMT-S proved to be acceptable (Cronbach's Alpha = .65 and .71; George & Mallery, 2002). This reliability fits to the results of Verhallen et al. (2017), who reported a Cronbach's Alpha of .71. Together with the high item difficulties (very easy items), the GFMT-S should be evaluated as too easy for psychological assessment of superior face recognition ability (standards see American Educational Research Association, 2014). Therefore, the application and interpretation of the norms for the GFMT-S for SR classification should be considered as preliminary and interpreted with caution. According to the request of Ramon et al. (2019), it would be promising to apply real-world face matching tests (using realistic face stimuli) in a multi-trait-multi-method approach including the CFMT+.

3.3.3 | Limitations and future directions

To summarize, a strength of the present study is the large online sample allowing to perform T norms (see Supporting information). The present data reveal the first systematic experimental investigation of presentation mode effects on the CFMT+, psychometric properties and the construct validity of the CFMT+ as a diagnostic tool for SR classification in German samples. However, the norm for the laboratory CFMT+ should be extended to include a more representative distribution of age and a more balanced gender ratio. It should also be mentioned that Sample 2 (online) was not selected according to face recognition ability and that the participants were not recruited in a random selection. However, it is possible that participants had differential motivations to perform the lab versus online CFMT+. Future research should investigate the reasons for the better online CFMT+ performance (effects of sample size could be excluded: results remain robust when randomly selected samples with the same group size were obtained from Sample 2) and examine how this affects the application of norms.

Further ideas for future research refer to the evaluation of the quality of face recognition tests with item-response-models (IRT models, cf. Cho et al., 2015, for the CFMT). Moreover, future research could use other analytical approaches to analyze the test performance of subgroups (bootstrap sampling or yoked sampling). With age-appropriate norms, the CFMT+ could be used for the entire age spectrum. Adapting the CFMT+ images to the age group of the participants could also helpful to better understand the development of the face recognition ability.

3.4 | Conclusion

Taken together, this study highlights the importance of examining the psychometric quality of the CFMT+ and the GFMT-S in different presentation modes. As one of the first, we provide T norms for the CFMT+ and GFMT-S in German samples. We show that gender and age of participants modulate face recognition performance in CFMT+, but not in GFMT-S. This opens further questions on whether CFMT+ performance results in criterion differences and whether norms for subgroups depending on age or gender should be applied. Moreover, it should be investigated whether performance differences for presentation mode affect criterion differences and, thus, suggest separate test norms even for presentation mode.

ACKNOWLEDGMENT

L.A.P. und A.L. designed the study. L.A.P. collected and analyzed the data. We are very grateful to Sabrina Voß and Eva Hildebrandt for their assistance during data collection. L.A.P. wrote the manuscript with A.L. This study is part of L.A.P.'s doctoral thesis.

CONFLICT OF INTEREST

We confirm that we have no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author, LAP, upon reasonable request.

ORCID

Lara Aylin Petersen https://orcid.org/0000-0002-1489-0943
Anja Leue https://orcid.org/0000-0002-2588-5226

REFERENCES

American Educational Research Association; American Psychological Association; National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., Wills, H., & Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications*, *3*, 22. https://doi.org/10.1186/s41235-018-0116-5

Belanova, E., Davis, J. P., & Thompson, T. (2018). Cognitive and neural markers of super-recognisers' face processing superiority and enhanced cross-age effect. *Cortex; A Journal Devoted to the Study of the Nervous System and Behavior*, *108*, 92–111. https://doi.org/10.1016/j.cortex.2018.07.008

Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *82*, 48–62. https://doi.org/10.1016/j.cortex.2016.05.003

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PLoS One*, *11*(2), e0148148. https://doi.org/10.1371/journal.pone.0148148

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

Bobak, A. K., Mileva, V. R., & Hancock, P. J. (2018). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology*, *72*(4), 872–881. https://doi.org/10.1177/1747021818776145

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of Young British adults. *Frontiers in Psychology*, *7*, 1378. https://doi.org/10.3389/fpsyg.2016.01378

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., Rivolta, D., Wilson, C. E., & Yovel, G. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant-stimulus ethnic match on the Cambridge face memory test and Cambridge face perception test. *Cognitive Neuropsychology*, *26*(5), 423–455. https://doi.org/10.1080/02643290903343149

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion [Introduction to test and questionnaire construction] (3., aktualisierte und erw. Aufl.)*. München: Pearson Studium.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, *42*(1), 286–291. https://doi.org/10.3758/BRM.42.1.286

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Cho, S.-J., Wilmer, J., Herzmann, G., McGugin, R. W., Fiset, D., van Gulick, A. E., Ryan, K. F., & Gauthier, I. (2015). Item response theory analyses of the Cambridge face memory test (CFMT). *Psychological Assessment*, *27*(2), 552–566. https://doi.org/10.1037/pas0000068

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates. https://doi.org/10.4324/9780203771587.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Davis, J. P., Bretfelean, L. D., Belanova, E., & Thompson, T. (2020). Super-recognisers: Face recognition performance after variable delay intervals. *Applied Cognitive Psychology*, *45*(3), 363–1368. https://doi.org/10.1002/acp.3712

Davis, J. P., Forrest, C., Treml, F., & Jansari, A. (2018). Identification from CCTV: Assessing police super-recogniser ability to spot faces in a crowd and susceptibility to change blindness. *Applied Cognitive Psychology*, *32*(3), 337–353. https://doi.org/10.1002/acp.3405

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior face recognition ability in police super-recognisers. *Applied Cognitive Psychology*, *30*(6), 827–840. https://doi.org/10.1002/acp.3260

DeGutis, J., Mercado, R. J., Wilmer, J., & Rosenblatt, A. (2013). Individual differences in holistic processing predict the own-race advantage in recognition memory. *PLoS One*, *8*(4), e58253. https://doi.org/10.1371/journal.pone.0058253

Dennett, H. W., McKone, E., Tavashmi, R., Hall, A., Pidcock, M., Edwards, M., & Duchaine, B. (2012). The Cambridge car memory test: A task matched in format to the Cambridge face memory test, with norms, reliability, sex differences, dissociations from face memory, and expertise effects. *Behavior Research Methods*, *44*(2), 587–605. https://doi.org/10.3758/s13428-011-0160-2

Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology*, *106*(3), 433–445. https://doi.org/10.1111/bjop.12103

Duchaine, B., & Nakayama, K. (2006). The Cambridge face memory test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576–585. https://doi.org/10.1016/j.neuropsychologia.2005.07.001

Fysh, M. C., Stacchi, L., & Ramon, M. (2020). Differences between and within individuals, and subprocesses of face cognition: Implications for theory, research and personnel selection. *Royal Society Open Science*, *7*(9), 200233. https://doi.org/10.1098/rsos.200233

George, D., & Mallery, P. (2002). *SPSS for windows step by step: A simple guide and reference. 11.0 upgrate*. Boston: Allyn & Bacon.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857. https://doi.org/10.3758/s13423-012-0296-9

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition*, *21*(9-10), 1306–1336. https://doi.org/10.1080/13506285.2013.823140

Herzmann, G., Danthiir, V., Schacht, A., Sommer, W., & Wilhelm, O. (2008). Toward a comprehensive test battery for face cognition: Assessment of the tasks. *Behavior Research Methods*, *40*(3), 840–857. https://doi.org/10.3758/BRM.40.3.840

McCaffery, J. M., Robertson, D. J., Young, A. W., & Burton, A. M. (2018). Individual differences in face identity processing. *Cognitive Research: Principles and Implications*, *3*, 21. https://doi.org/10.1186/s41235-018-0112-9

McCall, W. A. (1939). *Measurement: A Revision of How to Measure Education*. Macmillan.

Noyes, E., Hill, M. Q., & O'Toole, A. J. (2018). Face recognition ability does not predict person identification performance: Using individual data in the interpretation of group results. *Cognitive research: principles and implications*, *3*(1), 23. https://doi.org/10.1186/s41235-018-0117-4

Noyes, E., & O'Toole, A. J. (2017). Face recognition assessments used in the study of super-recognisers, arXiv, preprint arXiv: 1705.04739.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory (3. Ed.). McGraw-Hill series in psychology*. New York: McGraw-Hill.

Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology (London, England: 1953)*, *110*(3), 461–479. https://doi.org/10.1111/bjop.12368

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: A meta-analytic and theoretical review. *Psychological Bulletin*, *138*(1), 146–174. https://doi.org/10.1037/a0025750

Robertson, D. J., & Bindemann, M. (2019). Consolidation, wider reflection, and policy: Response to 'Super-recognisers: From the lab to the world and back again'. *British Journal of Psychology (London, England: 1953)*, *110*(3), 489–491. https://doi.org/10.1111/bjop.12393

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by metropolitan police super-Recognisers. *PLoS One*, *11*(2), e0150036. https://doi.org/10.1371/journal.pone.0150036

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252–257. https://doi.org/10.3758/PBR.16.2.252

Satchell, L. P., Davis, J. P., Julle-Danière, E., Tupper, N., & Marshman, P. (2019). Recognising faces but not traits: Accurate personality judgment from faces is unrelated to superior face memory. *Journal of Research in Personality*, *79*, 49–58. https://doi.org/10.1016/j.jrp.2019.02.002

Soloman, S. R., & Sawilowsky, S. S. (2009). Impact of rank-based normalizing transformations on the accuracy of test scores. *Journal of Modern Applied Statistical Methods*, *8*(2), 448–462. https://doi.org/10.22237/jmasm/1257034080

Statistisches Bundesamt (Destatis) (2019). Statistisches Jahrbuch – Deutschland und Internationales. Retrieved from https://www.destatis.de/DE/Themen/Querschnitt/Jahrbuch/_inhalt.html

Susilo, T., Germine, L., & Duchaine, B. (2013). Face recognition ability matures late: Evidence from individual differences in young adults. *Journal of Experimental Psychology. Human Perception and Performance*, *39*(5), 1212–1217. https://doi.org/10.1037/a0033469

Verhallen, R. J., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Bargary, G., & Mollon, J. D. (2017). General and specific factors in the processing of faces. *Vision Research*, *141*, 217–227. https://doi.org/10.1016/j.visres.2016.12.014

Woerner, W., Müller, C., & Hasselhorn, M. (2017). Bedeutung und Berechnung der Prozentränge und T-Werte beim Erstellen von Testnormen: Anmerkungen und Empfehlungen. Begabungen und Talente [Meaning and calculation of percentile ranks and T-values when generating test standards: Remarks and recommendations. Gifts and talents]. *Jahrbuch der pädagogisch-psychologischen Diagnostik, Tests & Trends*, *15*, 245–263.

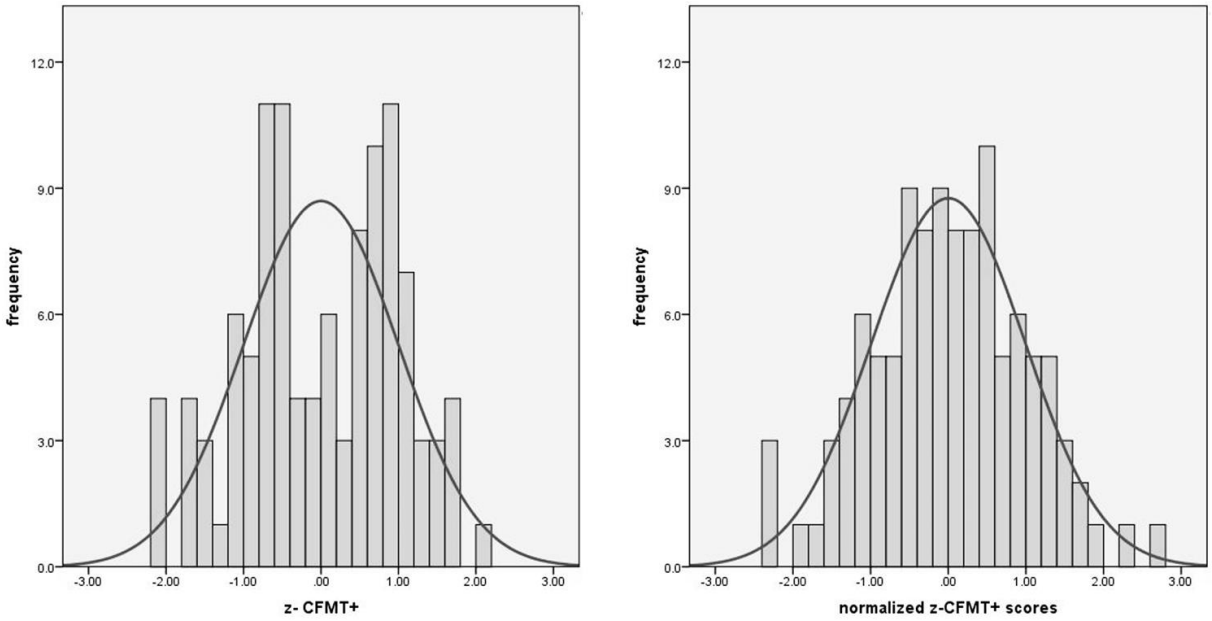
World Medical Association. (2013). Declaration of Helsinki: Ethical principles for medical research involving human subjects. *Journal of the American Medical Association*, *310*(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Petersen LA, Leue A. Extraordinary face recognition performance in laboratory and online testing. *Appl Cognit Psychol*. 2021;1–11. https://doi.org/10.1002/acp.3805

3.3 Anhang zur Studie 1 (Supplement)



A B

S1. Sample 1, laboratory. Distribution (kurtosis = -0.791, skewness = -0.170, *N* = 109) of the CFMT+ scores and the normal distribution curve. (A) z-transformed, non-normalized CFMT+ scores, (B) normalized z-scores of the CFMT+ with Rankit transformation (Solomon & Sawilowsky, 2009).

S2. Norms for CFMT+ scores, laboratory (using McCall transformation, McCall, 1939)

Raw score CFMT+ lab	**T score CFMT+ lab**
0–16*	20*–21*
17–32*	22*–23*
33–49*	24*–26*
50	27*–29
51–53*	30*–32
54*–55	33–34*
56	35–36*
57–58*	37

S2. (continued).

59–60*	38
61–62	39
63	40
64–65	41–42*
66	43
67	44–45*
68	46
69	47
70–71	48
72–73	49
74–76	50
77–78	51
79–80	52
81	53
82	54
83	55
84*	56
85	57–58*
86	59–60*
87	61
88	62
89–90	63
91	64
92	65
93	66–67*
94	68*–69
95	70*–71
96	72*–74
97–98*	75*
99	76*

S2. (continued).

100*	77–78*
101*	79*
102*	80*

*Note. Minimum raw score CFMT+ = 0, Maximum raw score CFMT+ = 102. * = raw scores or T score were extrapolated because they did not occur in the data.*

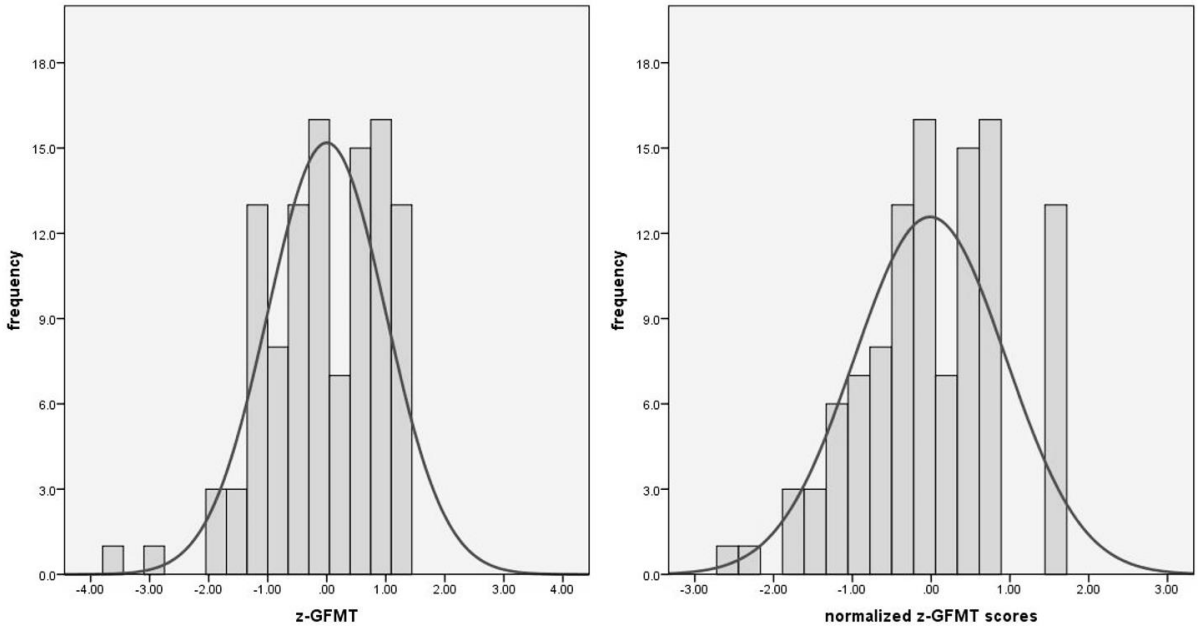
S3. Norms for CFMT+ scores, laboratory (using Rankit transformation, Solomon & Sawilowsky, 2009)

Raw Score CFMT+ lab	**T score for CFMT+ lab**
0*–49*	≤ 27*
50	28–30*
51	31
52*–54*	32
55	33
56	34–35*
57	36
58–59*	37
60*–61	38
62	39
63–64	40
65	41–42*
66	43
67	44
68	45
69	46
70	47
71–72	48
73–74	49
75–76	50
77–78	51
79–80	52

S3. (continued).

81	53
82	54
83–84	55–56*
85	57–58*
86	59–60*
87	61
88	62
89–90	63
91	64
92	65
93	66–67*
94	68–69*
95	70–71*
96	72
97*	73–74*
98*	75*
99	76
100*	77–78*
101*	79*
102*	80*

*Note. Minimum raw score CFMT+ = 0, Maximum raw score CFMT+ = 102. * = raw scores or T score were extrapolated because they did not occur in the data.*



A B

S4. Sample 1, laboratory. Distribution (kurtosis = 0.85, skewness = -0.87) of the GFMT-S scores and the normal distribution curve. (A) z-transformed, non-normalized GFMT-S scores, (B) normalized z-scores of the GFMT-S with Rankit transformation (Solomon & Sawilowsky, 2009).

S5. Norms for GFMT-S scores, laboratory (using McCall transformation, McCall, 1939)

Raw Score GFMT–S lab	**T score for GFMT–S lab**
0–8*	20*
9–18*	22*
19–24*	24*–25*
25	26
26*	27*–28*
27	29
28–29*	31*
30	32
31	35
32	37
33	40

S5. (continued).

34	43
35	46
36	49
37	52
38	54
39	59
40	66
40*	67*–80*

*Note. Minimum raw score GFMT-S = 0, Maximum raw score GFMT-S = 40. * = raw scores and T scores did not occur in the data. +3SD of the T-norm could not be computed because the data show a tendency of a two-peak distribution so that McCall transformation compensates only partly for deviation from normal distribution. The T-norm for the GFMT-S lab version is a preliminary estimation.*

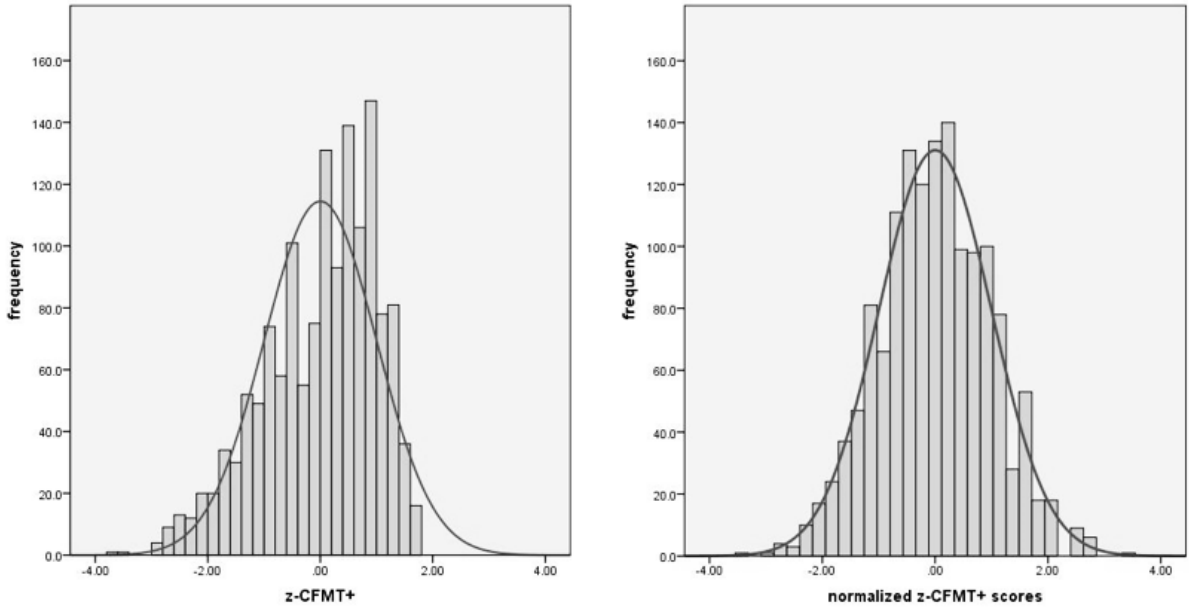
S6. Norms for GFMT-S scores, laboratory (using Rankit transformation, Solomon & Sawilowsky, 2009)

Raw Score GFMT–S – Lab	**T score for GFMT–S**
0*–8*	20
9*–18*	21*
19*–24*	22*–23*
25	24–25*
26*	26*
27	27*–28
28*	29*
29*	30*
30	31–32*
31	33*–34
32	35*–37
33	38*–40
34	41*–43
35	44*–45
36	46*–49

S6. (continued).

37	50*–52
38	53*–54
39	55*–59
40	60*–66
40*	67*–80*

*Note. Minimum raw score GFMT-S = 0, Maximum raw score GFMT-S = 40. * = raw scores or T score were extrapolated because they did not occur in the data. +3SD of the T-norm could not be computed because the data show a tendency of a right-shifted distribution so that Rankit transformation compensates only partly for deviation from normal distribution. The T-norm for the GFMT-S lab version is a preliminary estimation.*



S7. Sample 2, online. Distribution (kurtosis = -0.14, skewness = -0.64) of the CFMT+ scores and the normal distribution curve. (A) z-transformed, non-normalized CFMT+ scores, (B) normalized z-scores of the CFMT+ with Rankit transformation (Solomon & Sawilowsky, 2009).

S8. Norms for CFMT+ scores, online (using McCall transformation, McCall, 1939)

Raw score CFMT+ online	**T score CFMT+ online**
0–10*	20*
11–22*	20*
23–34*	20*
35–37*	20*
38–41*	20–21*
42*–44	22–23*
45–46	24
47	25
48	26
49	27–28*
50	29
51	30
52–53*	31
54–55	32
56	33
57–58	34
59	35
60–61	36
62	37
63–64	38
65	39
66–67	40
68–69	41
70	42
71–72	43
73	44
74–75	45
76–77	46
78	47
79–80	48

S8. (continued).

81	49
82	50
83	51
84–85	52
86	53
87	54
88	55
89	56
90	57–58*
91	59
92	60
93	61
94	62–63*
95	64
96	65–66*
97	67–68*
98	69–70*
99	71–72*
100	73*–74
101	78–79*
102	80

*Note. Minimum raw score CFMT+ = 0, Maximum raw score CFMT+ = 102. * = raw scores or T score were extrapolated because they did not occur in the data.*

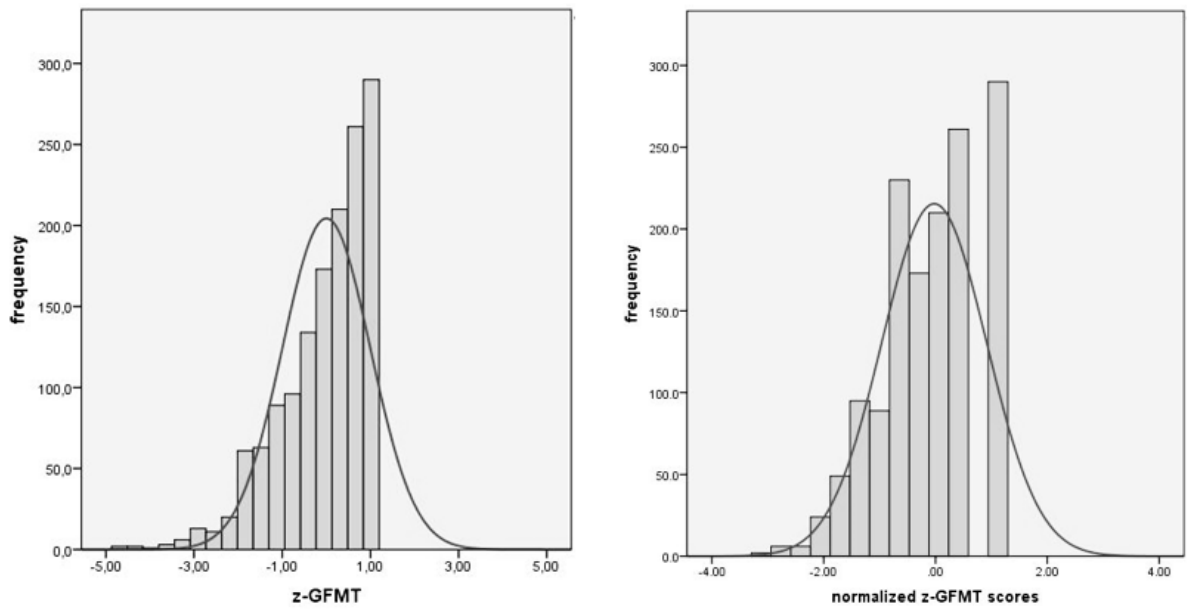
S9. Norms for CFMT+ scores, online (using Rankit transformation, Solomon & Sawilowsky, 2009)

Raw Score CFMT+ online	**T score for CFMT+ online**
0*–38	20
39–43*	21
44	22
45	23
46	24
47	25
48	26
49	27–28*
50	29
51	30
52–53	31
54–55	32
56	33
57–58	34
59	35
60–61	36
62	37
63–64	38
65–66	39
67	40
68–69	41
70	42
71–72	43
73	44
74–75	45
76–77	46
78	47
79–80	48
81	49

S9. (continued).

82	50
83	51
84–85	52
86	53
87	54
88	55
89	56
90	57–58*
91	59
92	60
93	61
94	62–63*
95	64
96	65–66*
97	67–68*
98	69–70*
99	71–72*
100	73*–74
101	75*–78
102	79*–80

*Note. Minimum raw score CFMT+ = 0, Maximum raw score CFMT+ = 102. * = raw scores or T score were extrapolated because they did not occur in the data.*



A B

S10. Sample 2, online. Distribution (kurtosis = 1.85, skewness = -1.31) of the GFMT-S scores and the normal distribution curve. (A) z-transformed, non-normalized GFMT scores, (B) normalized z-scores of the GFMT-S with Rankit transformation (Solomon & Sawilowsky, 2009).

S11. Norms for GFMT-S scores, online (using McCall transformation, McCall, 1939)

Raw Score GFMT–S online	**T score for GFMT–S online**
0–6*	20*
7–15*	20*
16–22*	20*
23	20*
24	21*–22
25	23
26	24
27	25*–26
28	27*–28
29	29*–30
30	31*–32

S11. (continued).

31	33*–34
32	35–36*
33	37–38*
34	39*–40
35	41*–42
36	43*–45
37	46*–48
38	49*–51
39	52*–55
40	56*–63
40*	64*–80*

*Note. Minimum raw score GFMT-S = 0, Maximum raw score GFMT-S = 40. * = raw scores did not occur in the data. +3SD of the T-norm could not be computed because the data show a tendency of a right-shifted distribution so that McCall transformation compensates only partly for deviation from normal distribution. The T-norm for the GFMT-S online version is a preliminary estimation.*

S12. Norms for GFMT-S scores, online (using Rankit transformation, Solomon & Sawilowsky, 2009)

Raw Score GFMT–S –Online	**T score for GFMT–S Online**
0*–23	20
24	21–22*
25	23
26	24–25*
27	26–27*
28	28–29*
29	30–31*
30	32–33*
31	34
32	35–36*
33	37–38*
34	39*–40

S12. (continued).

35	41*–42
36	43*–45
37	46*–48
38	49*–51
39	52*–55
40	56*–63
40*	64*–80*

*Note. Minimum raw score GFMT-S = 0, Maximum raw score GFMT-S = 40. * = raw scores did not occur in the data and were matched to the minimum T score. +3SD of the T-norm could not be computed because the data show a tendency of a right-shifted distribution so that Rankit transformation compensates only partly for deviation from normal distribution. The T-norm for the GFMT-S online version is a preliminary estimation.*

4 Studie 2 - Face memory and face matching: Internal consistency and test-retest reliability for the CFMT+ and the GFMT-S

Petersen, L. A., & Leue, A. (2022). Face memory and face matching: Internal consistency and test-retest reliability for the CFMT+ and the GFMT-S. *Journal of Individual Differences.* Advance online publication http://dx.doi.org/10.1027/1614-0001/a000361

In diesem Kapitel wird Studie 2 vorgestellt, die unter dem Titel „*Face memory and face matching: Internal consistency and test-retest reliability for the CFMT+ and the GFMT-S*“ in der Fachzeitschrift „*Journal of Individual Differences*“ (*Impact Factor* laut Homepage der Fachzeitschrift vom 10.09.2021: 2.08) veröffentlicht wurde. Nachfolgend werden in Abschnitt 4.1 der theoretische Hintergrund, die Methodik und die Ergebnisse zusammenfassend dargestellt sowie einige Diskussionspunkte aus dem Artikel angesprochen. Das englischsprachige, zur Publikation akzeptierte Manuskript zur Studie 2 ist in Abschnitt 4.2 zu finden, sodass dort detaillierte Informationen zu den hier zusammengefassten Inhalten nachgelesen werden können.

4.1 Zusammenfassung Studie 2

Theoretischer Hintergrund

Obwohl der CFMT+ (Russell et al., 2009) und der GFMT-S (Burton et al., 2010) häufig in der Gesichtserkennungsforschung eingesetzt werden (Bate et al., 2021; Ramon et al., 2019), gibt es kaum Berichte über die psychometrische Qualität der Tests. Dabei ist es gerade zur Klassifizierung von Leistungsunterschieden in der Gesichtserkennung wichtig, dass die Testergebnisse reliabel gemessen werden (vgl. American Educational Research Association et al., 2014). Die Reliabilität kann unter anderem durch die Untersuchung der internen Konsistenz eines Tests und die Test-Retest-Korrelation bewertet werden (vgl. American Educational Research Association et al., 2014). Zur Test-Retest-Reliabilität des CFMT+ gibt es noch keine

Befunde und für den GFMT-S liegt nur eine Publikation vor (Stantic et al., 2021). Die Test-Retest-Reliabilität ist wichtig, um Auskunft über die Stabilität einer Messung zu erhalten (Cronbach, 1947). Die Gesichtserkennungsfähigkeit sollte ein transsituationales, konsistentes und stabiles Konstrukt (*trait*) sein, da sie genetisch bedingt und weitgehend nicht trainierbar ist (Wilmer, 2017). Wenn die Gesichtserkennung eine individuelle, stabile Fähigkeit ist, sollte die interne Konsistenz eines Tests zur Messung des Konstrukts hoch sein. Als Index der internen Konsistenz und der Itemhomogenität integriert Cronbach's Alpha die mittleren Inter-Item-Korrelationen (MIC; Bernardi, 1994). Je höher die MICs des CFMT+ und des GFMT-S sind, desto intern konsistenter sollten beide Tests zur Gesichtserkennung sein. Tests, die intern konsistent sind und stabile Konstrukte messen, sollten auch mit einer höheren Test-Retest-Reliabilität einhergehen (Gregory, 2014). Neben der Berechnung der Test-Retest-Reliabilität kann man darüber hinaus in einem Studiendesign mit Testwiederholung im Sinne der Konstruktvalidität die Mittelwertsunterschiede zwischen zwei Messzeitpunkten untersuchen (Cronbach & Meehl, 1955). Es ist möglich, dass ein Test neben dem intendierten Konstrukt zu einem Messzeitpunkt noch etwas anderes misst, zum Beispiel Übungs- oder Erinnerungseffekte, und so das intendierte Konstrukt nicht allein gemessen wird. Zusammenfassend wurden in dieser Studie folgende vier Forschungsfragen untersucht:

1. Wie ist die Test-Retest-Reliabilität des CFMT+ im Zusammenhang mit der internen Konsistenz zu bewerten?
2. Wie ist die Test-Retest-Reliabilität des GFMT-S im Zusammenhang mit der internen Konsistenz zu bewerten?
3. Wie sind die Mittelwertsunterschiede zweier Messzeitpunkte im CFMT+ im Kontext der Konstruktvalidität zu bewerten?
4. Wie sind die Mittelwertsunterschiede zweier Messzeitpunkte im GFMT-S im Kontext der Konstruktvalidität zu bewerten?

Methodik

In einem Abstand von 12 Wochen absolvierten 72 Proband:innen (68.50% Frauen, $M_{\text{Alter}} = 44.38$ Jahre alt, 21 bis 69 Jahre alt) zweimal online den CFMT+ und den GFMT-S (Erhebungszeitraum im Jahr 2017). Alle Proband:innen kannten die Tests zum ersten Messzeitpunkt nicht und waren zum ersten Messzeitpunkt weitgehend Teil der Onlinestichprobe der Studie 1. Durch eine E-Mail wurden die Proband:innen von Studie 2 zur Teilnahme am zweiten Messzeitpunkt eingeladen. Es wurde wie in Studie 1 sichergestellt, dass die Daten die a-priori Kriterien für zuverlässige Online-Daten erfüllten. Die Tests wurden zu beiden Messzeitpunkten identisch in fester Reihenfolge durchgeführt, wieder über SoSci-Survey (siehe Methodik in Abschnitt 3.1): zuerst der CFMT+, eine Minute Pause, dann der GFMT-S. Die Proband:innen bearbeiteten die Tests in der Regel zu Hause oder im Büro und wurden zum zweiten Messzeitpunkt aufgefordert, dieselben Durchführungsbedingungen wie zum ersten Messzeitpunkt herzustellen (selbe Uhrzeit und selber Ort). Zur Beurteilung der internen Konsistenz wurden für jeden Messzeitpunkt die Koeffizienten Cronbach's Alpha, die Split-Half-Reliabilität mit Spearman-Brown-Korrektur sowie die mittleren Inter-Item-Korrelationen (MIC) berechnet. Die Split-Half-Reliabilität wurde dabei jeweils nach zwei verschiedenen Methoden zur Testhalbierung berechnet: „Odd-Even-Methode" (jedes zweite Item in eine Testhälfte) und „Erste-Zweite-Testhälfte-Methode" (Test wird in der Mitte der Items in erste und zweite Hälfte geteilt).

Ergebnisse

Aufgrund der nicht vorhandenen Normalverteilung der Testergebnisse wurden zunächst Spearman-Rank-Korrelationen als Test-Retest-Koeffizienten für den CFMT+ (Summenwert) und den GFMT-S (Summenwert) berechnet. Die Test-Retest-Reliabilität des CFMT+ kann mit $Rho(72) = .89$ ($p < .001$) als zufriedenstellend bewertet werden (vgl. Gregory, 2014). Alle Koeffizienten der internen Konsistenz waren für den CFMT+ größer als .87 und lagen

überwiegend im sehr guten Bereich über .90 (Bühner, 2011; George & Mallery, 2020). Die MIC als Maß der Itemhomogenität für den ersten Messzeitpunkt lag unter dem empfohlenen Bereich von .15 bis .50 (Clark & Watson, 2016), die MIC für den zweiten Messzeitpunkt lag innerhalb dieses Bereichs mit .18. Die Testwerte des zweiten Messzeitpunktes (M = 88.71, Md = 93.00, SD = 11.30) waren signifikant größer ($p < .001$; Wilcoxon-Vorzeichen-Rang-Test) als zum ersten Messzeitpunkt (M = 84.22, Md = 87.00, SD = 11.06).

Im Gegensatz zum CFMT+ kann die Test-Retest-Reliabilität des GFMT-S mit $Rho(72) = .68$ ($p < .001$) nicht als zufriedenstellend beurteilt werden, da dieser Wert unter .70 für einen zufriedenstellenden Koeffizienten liegt (vgl. Gregory, 2014; Souza, Alexandre, & Guirardello, 2017). Ebenso lagen alle Koeffizienten der internen Konsistenz bis auf eine Ausnahme unter .70. Daher kann die interne Konsistenz für den GFMT-S nicht als akzeptabel bewertet werden (vgl. Bühner, 2011; George & Mallery, 2020). Dazu passend lagen die MICs unter .15 und weisen damit auf Messfehler im GFMT-S hin. Die Testmittelwerte des GFMT-S waren zum zweiten Messzeitpunkt (M = 37.70, Md = 38.00, SD = 2.40) ebenfalls signifikant größer (p = .03, Wilcoxon-Vorzeichen-Rang-Test) als zum ersten Messzeitpunkt (M = 37.01, Md = 38.00, SD = 2.82). Bei der Interpretation dieses Ergebnisses sollten jedoch die geringen Reliabilitätskoeffizienten des GFMT-S berücksichtigt werden.

Diskussion

Insgesamt zeigte der CFMT+ in dieser Studie eine zufriedenstellende bis sehr gute Reliabilität (Test-Retest, Cronbach's Alpha und Split-Half-Koeffizienten; siehe Forschungsfrage 1). Nur die MICs zu beiden Messzeitpunkten waren niedrig, was auf mögliche Messfehler oder Einschränkungen der Itemhomogenität hinweist. Aufgrund der Art der Berechnung von Cronbach's Alpha kann es bei einer hohen Itemanzahl (102 Items im CFMT+) und niedrigen MIC dennoch möglich sein, dass Cronbach's Alpha einen hohen Wert erreicht (Bernardi, 1994). Es ist darüber hinaus möglich, dass die heterogenen Stimuli des CFMT+ die

niedrigen MICs erklären. So variieren die Gesichter der Stimuli in Perspektive und Bildqualität im CFMT+ (siehe Abbildung 1 in Abschnitt 1.4.1). Die Reliabilität des GFMT-S erwies sich in dieser Studie als nicht zufriedenstellend (Forschungsfrage 2). Die niedrigen Reliabilitätskoeffizienten weisen darauf hin, dass die Messung im GFMT-S Messfehlern unterliegen könnte, sodass Testwerte nur vorsichtig interpretiert werden sollten. Die geringen Reliabilitätskoeffizienten können jedoch nicht durch eine geringe Itemanzahl (40 Items) oder durch heterogene Stimuli erklärt werden. Die Stimuli des GFMT-S zeigen Gesichter in derselben Perspektive und Bildqualität, sodass sie hinsichtlich der Darstellungsform als homogen beschrieben werden können. Möglicherweise haben die niedrigen Itemvarianzen, die zum Teil Null waren (homogenes Antwortverhalten; Petersen & Leue, 2021), die Ausprägung von Cronbach's Alpha beeinflusst. So kann Cronbach's Alpha auch alternativ über die Varianzen der Items berechnet werden (Bernardi, 1994).

In dieser Studie wurden durch den Vergleich der Testmittelwerte zwischen den zwei Messzeitpunkten auch Hinweise auf die Konstruktvalidität des CFMT+ und des GFMT-S untersucht. Beide Tests zeigten signifikant höhere Testmittelwerte zum zweiten Messzeitpunkt. Beim CFMT+ lässt sich vermuten, dass zum zweiten Messzeitpunkt nicht nur das Konstrukt gemessen wurde, sondern die Messung durch Übungseffekte beeinflusst wurde (Forschungsfrage 3). Übungseffekte konnten bereits Murray und Bate (2020) für den kürzeren CFMT nachweisen. Daher wäre es möglich, dass Übungseffekte auch beim CFMT+ wirken. Beim GFMT-S könnten die Mittelwertsunterschiede (Forschungsfrage 4) jedoch auf die eingeschränkte Reliabilität oder auf Messfehler zurückzuführen sein (vgl. American Educational Research Association et al., 2014). Zusammenfassend hat diese Studie gezeigt, wie wichtig es ist, die Reliabilität von Gesichtserkennungstests wie dem CFMT+ und dem GFMT-S mit verschiedenen psychometrischen Koeffizienten integrativ zu bewerten, um Effekte von dem Stimulusmaterial oder dem Antwortverhalten auf die Reliabilität einschätzen zu können.

4.2 Englischsprachige Publikation der Studie 2

Es folgt das englischsprachige Manuskript zur Studie 2 mit dem Titel „*Face memory and face matching: Internal consistency and test-retest reliability for the CFMT+ and the GFMT-S*", das in der Fachzeitschrift „*Journal of Individual Differences*" veröffentlicht wurde. Das englischsprachige Manuskript enthält ein eigenes Literaturverzeichnis am Ende des Abschnitts.

Petersen, L. A., & Leue, A. (2022). Face memory and face matching: Internal consistency and test-retest reliability for the CFMT+ and the GFMT-S. *Journal of Individual Differences*. Advance online publication http://dx.doi.org/10.1027/1614-0001/a000361

Face memory and face matching: Internal consistency and test-retest reliability for the CFMT+ and the GFMT-S

Lara Aylin Petersen and Anja Leue

Kiel University

Abstract

The Cambridge Face Memory Test Long (CFMT+) and the Glasgow Face Matching Test Short (GFMT-S) are frequently used tests in face recognition research. No test-retest results in conjunction with internal consistency, mean inter-item correlations (MICs), and pre-post mean differences have been reported. The internal consistency and the MICs provide insights into the homogeneity of items. In an online study (N = 72), we investigated the test-retest reliability, Cronbach's Alpha, split-half reliability, MICs, and retest mean differences for the CFMT+ and the GFMT-S. The CFMT+ showed satisfactory reliability coefficients above .88, whereas the coefficients of the GFMT-S were mainly dissatisfactory and below .75. We argue that task characteristics like heterogeneous stimulus material might lower MICs, response behavior might enhance reliability, and practice effects might increase the means of the CFMT+ in repeated measurements. Therefore, an integrative evaluation of different psychometric parameters helps to explain variations of reliability in face recognition tests.

Words 150

Keywords: Test-retest reliability, internal consistency, face recognition, CFMT+, GFMT-S

Introduction

Face perception and face recognition are important components of social interaction and, for example, in forensic settings (Bruce & Young, 2012). Face memory and face matching abilities are in the focus of face recognition research (Bate et al., 2018; Bobak, Hancock & Bate, 2016; Ramon, Bobak & White, 2019) and face processing research (Verhallen et al., 2017). The Cambridge Face Memory Test Long Form (CFMT+) measuring short-term face memory (Russell, Duchaine & Nakayama, 2009) and the Glasgow Face Matching Test Short Form (GFMT-S) measuring face matching (Burton, White & Mc Neill, 2010) are frequently used in laboratory and online studies (Table 1 in Ramon et al., 2019). Currently, there are only a few reports of psychometric properties for the CFMT+ and the GFMT-S (e.g., Cronbach's Alpha in Petersen & Leue, 2021; Verhallen et al., 2017). To our knowledge, there are no test-retest reliability reports for the CFMT+ and only one for the GFMT-S (Stantic et al., 2021). According to the standards in psychological assessment (American Educational Research Association, 2014) it is important to report different psychometric properties. The magnitude of the reliability coefficients of the CFMT+ and the GFMT-S is important for applied and assessment settings like the personnel selection of police officers (cf. Ramon et al., 2019). The lower the reliability coefficients are, the less accurately tests measure a person's true score (American Educational Research Association, 2014). Test-retest reliability indicates how stable a construct can be measured over time and is operationalized as the correlation between test scores obtained for the same individuals on the same test at different measurement points (Cronbach, 1947). Due to daily fluctuations and other individual or situational conditions in a person's performance (e.g., concentration or fatigue), face recognition performance could differ across time (Busey & Loftus, 2007). Some studies suggest that face recognition performance should be highly test-retest reliable because face recognition is highly heritable (Shakeshaft & Plomin, 2015; Wilmer et al., 2010) and can be trained only slightly or not at all (Dolzycka, Herzmann,

Sommer & Wilhelm, 2014; Hillstrom, Sauer & Hope, 2011; White, Kemp, Jenkins, Matheson & Burton, 2014). Face recognition performance is important in eyewitness crime scenes (Bruce & Young, 2012). Crime scenes are often of very short observation times and occur in sub-optimal perception settings (Busey & Loftus, 2007) suggesting an interaction of heritable face recognition performance and requirements of the situation. This study presents test-retest reliability data for the Online-CFMT+ and the Online-GFMT-S in a context of low time pressure, low social pressure, and optimal perception settings.

If face recognition is a construct that describes a rather homogeneous individual ability Cronbach's Alpha coefficients should be high even in the assessment context of this study. Cronbach's Alpha integrates the mean inter-item correlations (MICs) and represents the conceptual coherence of all items [Formula: $\alpha = N*r\text{m} / 1 + (N - 1) * r\text{m}$; $r\text{m}$ = mean intercorrelation of N items] (Bernardi, 1994; Formula 2). Therefore, the higher the MICs in the CFMT+ and the GFMT-S, respectively, the more internally consistent the measured construct. Tests that are internally consistent and that assess stable constructs like traits should result in higher test-retest coefficients because measurement errors would less influence the test performance over time (Gregory, 2014).

In addition to test-retest reliability, mean differences of two measurement points can be interpreted in the context of construct validity (Cronbach & Meehl, 1955). Significant mean differences could indicate that memory or practice effects influence the performance in test repetitions suggesting higher test scores in the second measurement than in the initial measurement (Gregory, 2014; Lord & Novick, 2008). If practice effects would not drive the performance of the CFMT+ and the GFMT-S at different measurement time points, the mean values in the CFMT+ and the GFMT-S would not significantly differ over time. Consequently, if measurement errors like practice effects would not matter so that just an individual's heritable performance would drive test-retest scores of the CFMT+ and the GFMT-S, high test-retest

reliability coefficients should go along with non-significant mean value variations of the CFMT+ and the GFMT-S across measurement points. Therefore, in this study we wanted to analyze and to interpret the mean differences of two measurement points for the CFMT+ and the GFMT-S.

The CFMT+ (Russell et al., 2009) has been used in the research of superior face recognition (Bobak, Pampoulov & Bate, 2016; Davis, Lander, Evans & Jansari, 2016; Ramon et al., 2019). There are a few test-retest reliability reports for the CFMT (Duchaine & Nakayama, 2006), the easier former version of the CFMT+ (Murray & Bate, 2020; Stantic et al., 2021; Wilmer et al., 2010). Wilmer et al. (2010) reported a test-retest reliability of $r(389) = .70$ for a six months test-retest interval and $r(42) = .76$ for a two months test-retest interval (online CFMT). Thus, the test-retest reliability of the CFMT can be rated as low to satisfactory (Gregory, 2014). In Gregory (2014) a test-retest reliability of ≥ .70–.80 for performance tests is rated as satisfactory. The test-retest reliability for the GFMT-S was satisfactory (.77, 14 days, $N = 69$) in Stantic et al. (2021). Consistent with these results, the test-retest reliability has been reported for other face memory and face matching tests with low to satisfactory values, e.g., .59 (one week, $N = 78$; UNSW Face Test; Dunn, Summersby, Towler, Davis & White 2020) and .76 (one month, $N = 102$; Recognition Memory Test Faces; Bird, Papadopoulou, Ricciardelli, Rossor & Cipolotti, 2003). Overall, in this study we investigate the test-retest reliability in conjunction with the internal consistency, its MICs, and mean differences for repeated measures of the CFMT+ and the GFMT-S.

Methods

Participants

A total of $N = 75$ participants performed the CFMT+ and the GFMT-S online at two measurement points, 12 weeks apart according to the intervals in Wilmer et al. (2010). Three participants who did not fulfil the criteria for sufficient quality of online data were excluded.

For example, participants were excluded when they used a smartphone or reported technological problems. Similar criteria have been used in Petersen and Leue (2021). Thus, $N = 72$ participants (49 females, 68.06 %, $M = 44.38$ years, $SD = 12.40$, *range* 21–69 years) were included for data analysis. All participants were white Caucasian, lived in Germany, and performed the tests for the first time at the first measurement point. Regarding professional status, 84.72 % of the participants ($N = 61$) had a university degree or general qualification for university entrance. The participants were not selected according to their recognition ability.

Materials

The Cambridge Face Memory Test Long Form (CFMT+; Russell et al., 2009)

The CFMT+ is the more difficult version of the computer-based standardized CFMT (Duchaine & Nakayama, 2006) and measures short-term face memory with an identification task. The test familiarizes participants with six male target faces, gray-scaled, presented from three viewpoints. After the learning phase, three pictures were presented and the participant must choose which person is known from the learning phase, using three alternative forced-choice answer categories (for stimuli see Russell et al., 2009). The CFMT+ yields a total score of 102 points with one point for each correctly answered item. The CFMT+ has no time limit. Test properties, e.g., item difficulties and Cronbach's Alpha, of the adapted Online-CFMT+ are reported in Petersen and Leue (2021) for laboratory ($N = 109$) and online settings ($N = 1435$; Cronbach's Alpha = 0.92).

The Glasgow Face Matching Test Short Form (GFMT-S; Burton et al, 2010)

The short version of the GFMT (Burton et al., 2010) consists of 40 item pairs of simultaneously presented unfamiliar face images in a classical face matching task, half with same identity and half with different identity. The images show female and male faces in frontal position, with neutral expression, and in grayscale (for stimuli see Burton et al., 2010). Participants decided whether the faces are of the same person or of two different persons

(answer categories: same or different). The GFMT-S yields a total score of 40 points, one point per correctly answered item. The GFMT-S has no time limit. Some test properties, e.g., item difficulties and Cronbach's Alpha, of the adapted Online-GFMT-S conducted in laboratory (N = 109) and online settings (N = 1435; e.g., Cronbach's Alpha = .71) are reported in Petersen and Leue (2021).

Procedure

All data were collected online via SoSci Survey, a web-based platform for surveys (https://www.SoScisurvey.de/en/index). Out of a large online sample (N = 1435; Petersen & Leue, 2021) we recruited a smaller sub-sample for the present test-retest study. The study was designed according to the ethical principles of the Declaration of Helsinki (World Medical Association, 2013). The first and second measurements (12 weeks apart) followed the same procedure. The tests were presented in a fixed order: participants gave written informed consent, demographic questions, followed by the CFMT+, one-minute break, followed by the GFMT-S, control questions to operationalize the online data quality criteria (see Participants), and feedback on performance. The participants usually performed the tests at home or at work and were prompted to conduct the study in a quiet environment reducing disturbance effects, e.g., turn off mobile phone. At the second measurement point, participants were asked to perform the same test conditions as in the first measurement, e.g., same time and same place.

Statistical analysis

The CFMT+ and the GFMT-S performances were scored dichotomously for each item using the correct responses. Therefore, we could have calculated the internal consistency by using the Kuder-Richardson 20 formula. Because Cronbach's Alpha is identical with Kuder-Richardson 20 formula when items are scored 0 and 1 (Anselmi, Colledani & Robusto, 2019; Feldt, 1969), we report the Cronbach's Alpha coefficients. Bühner (2011) as well as George and Mallery (2020) interpreted reliability coefficients (e.g., Cronbach's Alpha) of < .80 to be

low, of .80–.90 to be moderate, and of > .90 to be high or excellent. For benchmarks on test-retest reliability, we followed the recommendations of Gregory (2014) and Souza et al. (2017) interpreting a test-retest reliability > .70 as satisfactory. Furthermore, we calculated the MICs and the split-half reliability. We used two methods to calculate the split-half reliability (both Spearman-Brown corrected): Odd-Even (relevant for difficulty graded tests like the CFMT+) and First-Second half method (relevant for applied contexts of eyewitness testimony; see Discussion). A Wilcoxon signed-rank test was performed to investigate mean differences for the two measurement points of the CFMT+ and the GFMT-S. All analyses were conducted with IBM SPSS Statistics (Version 26) for Windows.

Results

CFMT+

Kolmogorov-Smirnov (K-S) tests showed no normal distribution of the CFMT+ scores for the first and second measurement point (First: $K\text{-}S(72) = 0.15$, $p < .001$; Second: $K\text{-}S(72) = 0.21$, $p < .001$). Therefore, we calculated the Spearman Rank correlation for the test-retest reliability, $Rho(72) = .89$ ($p < .001$, two-tailed), which can be evaluated as satisfactory (Gregory, 2014). We calculated Cronbach's Alpha and the two split-half reliability coefficients for the first and second measurement point (Table 1). All coefficients are above .90, except the First-Second half method split-half reliability, r_{tt} = .87–.88 (Table 1). Therefore, the reliability can be mainly rated as excellent (George & Mallery, 2020). Furthermore, we calculated the MICs (Table 1). The MIC for the first measurement point (MIC = .12) is below the recommended range of .15–.50 (Clark & Watson, 2016), while the MIC for the second measurement can be evaluated as sufficient. The MICs suggest that error variance influenced the measurement of CFMT+ performance at the first measurement point.

Table 1

Cronbach's Alpha (with CI), MICs and split-half reliabilities of the CFMT+ and GFMT-S for the first and second measurement points (12 weeks apart).

	CFMT+ First time	CFMT+ Second time	GFMT-S First time	GFMT-S Second time
Cronbach's Alpha	.91	.94	.68	.64
CI	[.88; .94]	[.92; .96]	[.57; .78]	[.52; .76]
MIC	.12	.18	.07	.06
Split-half, method 1[a]	.91	.91	.65	.75
Split-half, method 2[b]	.87	.88	.68	.60

Notes. $N = 72$; CFMT+ = Cambridge Face Memory Test Long (102 items); GFMT-S = Glasgow Face Matching Test Short (40 items); CI = confidence interval for Cronbach's Alpha; MIC = mean inter-item correlation.

[a] Split-half reliability using Odd-Even method, Spearman-Brown corrected.

[b] Split-half reliability using First-Second half method, Spearman-Brown corrected.

In addition to reliability, we analyzed the mean differences between the repeated measurements with a Wilcoxon signed-rank test. The mean score for the first measurement ($M = 84.22$, *median* = 87.00, $SD = 11.06$, *range* = 52–101, skewness = -0.87) was significantly lower than the CFMT+ performance for the second measurement ($M = 88.71$, *median* = 93.00, $SD = 11.30$, *range* = 50–101, skewness = -1.52), $Z = -6.28$, $p < .001$, suggesting that contextual effects (e.g., practice or other measurement errors) influenced CFMT+ performance between both measurement points. According to George and Mallery (2020), a skewness between -1.00 and 1.00 is excellent. A negative skewness indicates a greater number of higher test scores (i.e., more correct responses) and therefore a homogeneous response behavior (Figure 1A).

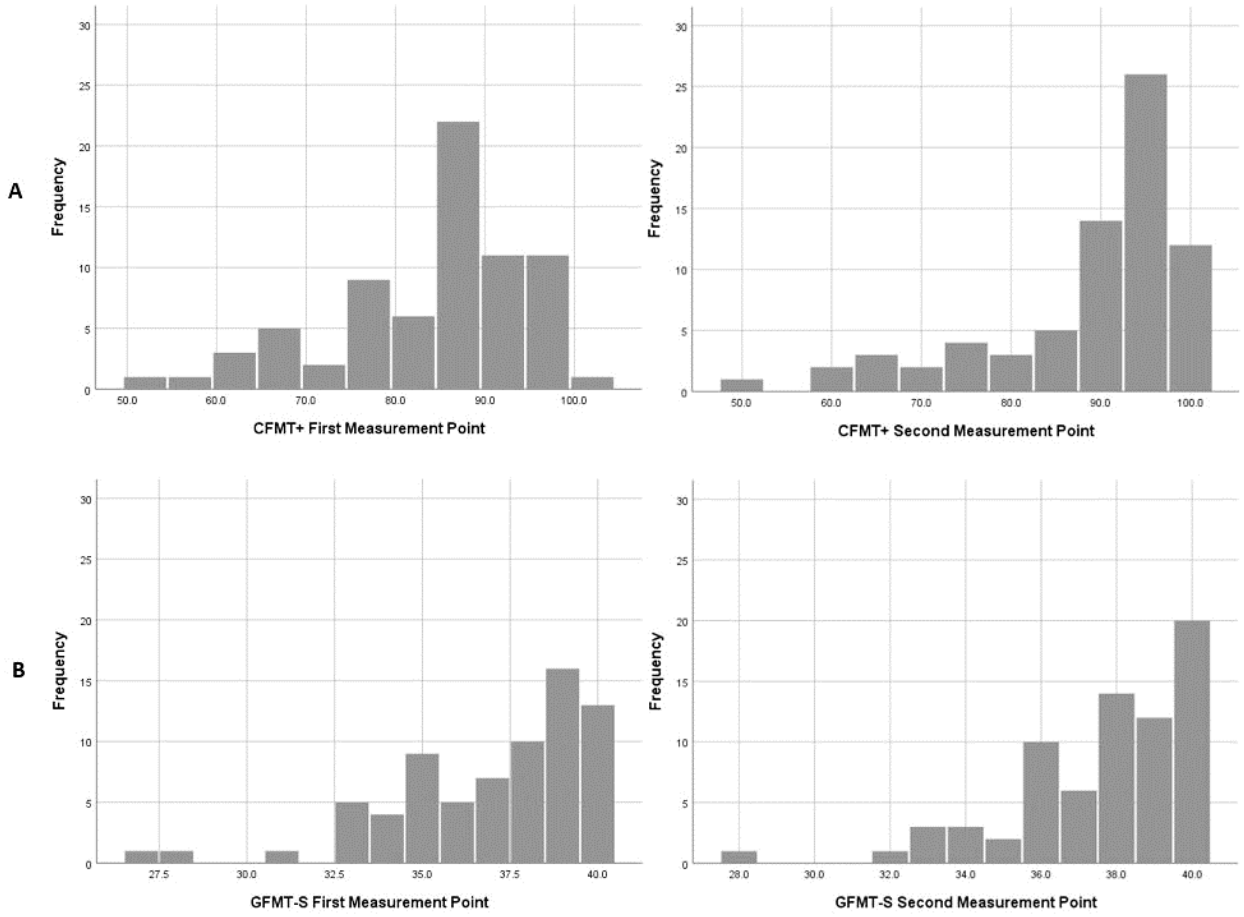


Figure 1. Distribution of the test scores in the CFMT+ (A) and in the GFMT-S (B) for the first measurement point (left) and the second measurement point (right).

GFMT-S

Kolmogorov-Smirnov (K-S) tests showed no normal distribution of the GFMT-S scores for the first and second measurement point (First: $K\text{-}S(72) = 0.18$, $p < .001$; Second: $K\text{-}S(72) = 0.19$, $p \leq .001$). Therefore, we calculated the Spearman Rank correlation for the test-retest reliability, $Rho(72) = .68$ ($p < .001$, two-tailed), which can be evaluated as low (Gregory, 2014). We also computed Cronbach's Alpha and the two split-half reliability coefficients (see Table 1) for the first and second measurement point. All internal consistency coefficients are below .70, except the Odd-Even split-half coefficient from the second measurement point with r_{tt} =.75 (Table 1). Therefore, the internal consistency can be evaluated

as low or at best acceptable (George & Mallery, 2020). The MICs for both measurement points (Table 1) were below the recommended range of .15–.50. (Clark & Watson, 2016) and suggest that error variance influenced the measurement of GFMT-S performance at both measurement points.

In addition to reliability, we analyzed the mean differences between the repeated measurements with a Wilcoxon signed-rank test. The mean score for the first measurement (*M* = 37.01, *median* = 38.00, *SD* = 2.82, *range* = 27–40, skewness = -1.28) was significantly lower than the mean of the second measurement (*M* = 37.70, *median* = 38.00, *SD* = 2.40, *range* = 28–40, skewness = -1.43), *Z* = -2.12, *p* = .03. According to George and Mallery (2020), the skewness of the distribution of the GFMT-S scores cannot be evaluated as excellent. Many participants reached a higher test score (Figure 1B).

Discussion

This study investigated the test-retest reliability, the internal consistency (Cronbach's Alpha and split-half reliability), and the MICs of the CFMT+ (Russell et al., 2009) and the GFMT-S (Burton et al., 2010). Further, we analyzed the repeated measures mean differences in terms of construct validity (Cronbach & Meehl, 1955). Almost all reliability coefficients for the CFMT+ were satisfactory to excellent, only the MICs were small. In contrast, the reliability coefficients and the MICs for the GFMT-S cannot be rated as satisfactory. In terms of construct validity, both tests showed significantly higher test performances at the second measurement point.

An integrative interpretation of the reliability with low MICs

The results showed a satisfactory test-retest reliability for the CFMT+ with values above .70 (Gregory, 2014; Souza et al., 2017). The test-retest reliability of the CFMT+ (*Rho*(72) = .89) is higher than the reported test-retest coefficient for the shorter CFMT (Duchaine & Nakayama, 2006) in Wilmer et al. (2010), e.g., *r*(42) = .76 for a two months interval. Further, the CFMT+

has proven to be reliable because Cronbach's Alpha and the split-half reliability coefficients were excellent for both measurement points with values above .90 (George & Mallery, 2020). Only the First-Second half method split-half reliability was slightly below .90. An increasing item difficulty is more balanced in the Odd-Even method compared to First-Second half method because both test halves in the Odd-Even method incorporate items with a continuously increasing item difficulty. Therefore, split-half reliability might be slightly higher for Odd-Even than for First-Second half method (see Table 1). First-Second half method reliability coefficients are also of interest in forensic settings. They provide information on how reliable line-ups in identification tasks might be when eyewitnesses would identify a suspect in the first half (i.e., before they have seen all suspects) or in the second half of a line-up. Moreover, a random selection of faces in the CFMT+ or GFMT+S for calculating reliability might be promising for a lower-bound calculation of reliability. Such a random selection of faces would be worthwhile as line-ups also apply varying stimuli for different suspects (chapter 6; Bruce & Young, 2012). Concerning the CFMT+, the high First-Second half method split-half reliability could be used to make performance predictions when a test taker finishes the test after the first half of the items, e.g., because of limited time or disruption. Items of the CFMT+ or the GFMT-S could be pre-selected based on values of both types of split-half reliability to exclude measurement errors that are due to sequence effects of the items, e.g., effects of continuously increasing item difficulty.

We further calculated the MICs as an index of item homogeneity (i.e., stimulus or perception of stimuli) because Cronbach's Alpha is influenced by the number of items and their inter-item correlations in a test (Bernardi, 1994; Formula 2). The CFMT+ has 102 items. Therefore, Cronbach's Alpha can be high despite low MICs just because of a large number of items. The MIC for the first measurement point was below the optimal reference range of Clark and Watson (2016). Since moderate to high inter-item correlations can be expected for good

reliability values, the low MICs indicate that the items of the CFMT+ incorporate measurement errors. One explanation for low MICs could be the effect of heterogeneous stimulus material. In each of the four sections of the CFMT+ the image material is modified to increase the difficulty of the items (see example images for the sections in Russell et al., 2009). Furthermore, faces from three different viewpoints are used as stimuli. Future research should investigate which errors (e.g., memory, stimulus material; Lord & Novick, 2008) might affect variations of Cronbach's Alpha coefficient and whether different types of measurement errors might influence the construct validity of the CFMT+. Unreliability is not necessarily due to the same types of error. Therefore, future research might systematize different types of non-random or random errors (Beauducel & Leue, 2014). For example, homogeneous response behavior (i.e., low variance of items) might be an explanation for a high Cronbach's Alpha coefficient even when the MICs are possibly low due to a heterogeneous stimulus material. Moreover, McDonald's Omega could be compared to Cronbach's Alpha in further studies (Hayes & Coutts, 2020). One difference between Cronbach's Alpha and Omega is that an essential tau equivalence is no prerequisite for Omega, but it is for Cronbach's Alpha (Hayes & Coutts, 2020). Therefore, calculating Cronbach's Alpha is critically discussed in Hayes and Coutts (2020).

In contrast to the CFMT+, the test-retest reliability of the GFMT-S underscores the recommended .70 (Gregory, 2014; Souza et al., 2017). Furthermore, both Cronbach's Alpha and the split-half reliability coefficients of the GFMT-S for the first and second measurement point were below .90 (George & Mallery, 2020). Thus, the reliability of the GFMT-S was not satisfying. Moreover, the MICs were small and below the reference range of .15–.50 (Clark & Watson, 2016). This indicates that a high proportion of variance in the GFMT-S scores is attributable to measurement errors. Therefore, test scores of the GFMT-S should be interpreted with caution. If the GFMT-S with 40 items had 102 items like the CFMT+, Cronbach's Alpha

would reach .84 (cf. Spearman-Brown prophecy formula assuming 2.55 as many items as currently given in the GFMT-S). Moreover, the low MICs for the GFMT-S cannot be easily explained because the stimulus material is more homogeneous than in the CFMT+. Possibly, the low item difficulties (see Petersen & Leue, 2021) distort the variance of the items because some items had a variance of 0. The reported test-retest reliability of the GFMT-S corresponds to values of other face matching tests like the KENT face matching test with a test-retest reliability of $r(28) = 0.67$ with a seven day interval (Fysh & Bindemann, 2018). Future research may revise or develop (new) face matching tests because the GFMT-S was not sufficiently (test-retest) reliable.

Construct validity: Mean Differences and practice effects

In this study, participants achieved on average a higher score in the CFMT+ in the second measurement ($M_2 = 88.71$) than in the first measurement ($M_1 = 84.22$). It is possible that a test measures different parts of a construct at different measurement points. In this line, it could be presumed that a test does not exclusively measure the intended construct at the second measurement point, but rather measures the construct plus an unsystematic or even systematic error, e.g., memory or practice effects (Lord & Novick, 2008). Since face memory has been considered a stable construct (e.g., Wilmer et al., 2010), intra-individual score variations in the CFMT+ should not be traced back to a trait change. Therefore, measurement errors seem to affect the construct validity of the CFMT+. This corresponds to Murray and Bate (2020) who reported practice effects for some sections of the CFMT (Duchaine & Nakayama, 2006). Future research could vary the interval between the measurement points or compare test-retest mean results of the same and different test versions to disentangle practice or memory effects on CFMT+ performance. For the GFMT-S, the mean difference between the measurement points were also significant but should be interpreted with caution ($M_1 = 37.01$, $M_2 = 37.70$) because the test was not sufficiently reliable (see above). Therefore, mean score variations of the

GFMT-S may be attributed to measurement errors or reliability restrictions (American Educational Research Association, 2014; Lord & Novick, 2018).

Conclusions

In conclusion, the CFMT+ showed a satisfactory test-retest reliability and a high to excellent internal consistency (Cronbach's Alpha and split-half reliability). However, the low MIC for the first measurement point indicates that the test performance might be influenced by errors (e.g., heterogeneous stimulus material). The significant mean difference between the two measurement points suggests the influence of practice effects for the CFMT+. The GFMT-S showed a limited internal consistency and not satisfying MICs. Therefore, the test scores could be influenced by measurement errors and the low test-retest reliability of the GFMT- S should be interpreted with caution. We recommend evaluating the reliability of face memory and face matching tests by calculating different psychometric coefficients (i.e., internal consistency, MICs, and test-retest reliability). This integrative evaluation of psychometric parameters of face recognition tests allows elucidating effects of stimulus material and response behavior on reliability.

References

American Educational Research Association; American Psychological Association; National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anselmi, P., Colledani, D., & Robusto, E. (2019). A Comparison of classical and modern measures of internal consistency. *Frontiers in Psychology*, *10*, 2714. https://doi.org/10.3389/fpsyg.2019.02714

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., . . . Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications*, *3*, 22. https://doi.org/10.1186/s41235-018-0116-5

Bernardi, R. A. (1994). Validating research results when Cronbach's Alpha is below .70. A Methodological Procedure. *Educational and Psychological Measurement, 54* (3), 766–775. https://doi.org/10.1177/0013164494054003023

Beauducel, A., & Leue, A. (2014). Testing the assumption of uncorrelated errors for short scales by means of structural equation modeling. *Journal of Individual Differences, 35*, 201-211. https://doi.org/10.1027/1614-0001/a000135

Bird, C. M., Papadopoulou, K., Ricciardelli, P., Rossor, M. N., & Cipolotti, L. (2003). Test-retest reliability, practice effects and reliable change indices for the recognition memory test. *The British Journal of Clinical Psychology*, *42*(Pt 4), 407–425. https://doi.org/10.1348/014466503322528946

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

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of young British adults. *Frontiers in Psychology*, *7*, 1378. https://doi.org/10.3389/fpsyg.2016.01378

Bruce, V., & Young, A. (2012). *Face perception*. London: Psychology Press.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*(1), 286–291. https://doi.org/10.3758/BRM.42.1.286

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte und erw. Aufl.) *[Introduction to test and questionnaire construction]*. *PS Psychologie*. München: Pearson Studium

Busey, T. A., & Loftus, G. R. (2007). Cognitive science and the law. *Trends in Cognitive Sciences*, *11*(3), 111–117. https://doi.org/10.1016/j.tics.2006.12.004

Clark, L. A., & Watson, D. (2016). Constructing validity: Basic issues in objective scale development. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research (4th ed.)* (pp. 187–203). Washington: American Psychological Association. https://doi.org/10.1037/14805-012

Cronbach, L. J. (1947). Test reliability; its meaning and determination. *Psychometrika*, *12*(1), 1-16. https://doi.org/10.1007/BF02289289

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior face recognition ability in police super-recognisers. *Applied cognitive psychology*, *30*(6), 827–840. https://doi.org/10.1002/acp.3260

Dolzycka, D., Herzmann, G., Sommer, W., & Wilhelm, O. (2014). Can training enhance face cognition abilities in middle-aged adults? *PloS One*, *9*(3), e90249. https://doi.org/10.1371/journal.pone.0090249

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576–585. https://doi.org/10.1016/j.neuropsychologia.2005.07.001

Dunn, J. D., Summersby, S., Towler, A., Davis, J. P., & White, D. (2020). UNSW Face Test: A screening tool for super-recognizers. *PloS one*, *15*(11), e0241747. https://doi.org/10.1371/journal.pone.0241747

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika, 34* (3), 363–373. https://doi.org/10.1007/BF02289364

Fysh, M. C., & Bindemann, M. (2018). The Kent Face Matching Test. *British Journal of Psychology*, *109*(2), 219–231. https://doi.org/10.1111/bjop.12260

George, D., & Mallery, P. (2020). *IBM SPSS statistics 26 step by step: A simple guide and reference* (Sixteenth edition). New York, London: Routledge.

Gregory, R. J. (2014). *Psychological testing: History, principles and applications* (7. ed., global ed.). Boston: Pearson Education.

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Communication Methods and Measures, 14*(1), 1-24. https://doi.org/10.1080/19312458.2020.1718629

Hillstrom, A. P., Sauer, J., & Hope, L. (2011). *Training methods for facial image comparison: A literature review*. The Stationary Office. Retrieved from https://eprints.soton.ac.uk/371613/

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. *Addison-Wesley series in behavioral science*. Charlotte, N.C.: Information Age Pub.

Murray, E., & Bate, S. (2020). Diagnosing developmental prosopagnosia. Repeat assessment using the Cambridge Face Memory Test. *Royal Society open science, 7* (9), 200884. https://doi.org/10.1098/rsos.200884

Petersen, L. A., & Leue, A. (2021). Extraordinary face recognition performance in laboratory and online testing. *Applied Cognitive Psychology*, *35*, 579-589. https://doi.org/10.1002/acp.3805

Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology, 110*(3), 461–479. https://doi.org/10.1111/bjop.12368

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review, 16*(2), 252–257. https://doi.org/10.3758/PBR.16.2.252

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(41), 12887–12892. https://doi.org/10.1073/pnas.1421881112

Souza, A. C. de, Alexandre, N. M. C., & Guirardello, E. de B. (2017). Psychometric properties in instruments evaluation of reliability and validity [Portuguese: Propriedades psicométricas na avaliação de instrumentos: Avaliação da confiabilidade e da validade]. *Epidemiologia E Servicos De Saude: Revista Do Sistema Unico De Saude Do Brasil*, *26*(3), 649–659. https://doi.org/10.5123/S1679-49742017000300022

Stantic, M., Brewer, R., Duchaine, B., Banissy, M. J., Bate, S., Susilo, T., . . . Bird, G. (2021). The Oxford Face Matching Test: A non-biased test of the full range of individual differences in face perception. *Behavior research methods*, *49*(9), 2541. https://doi.org/10.3758/s13428-021-01609-2

Verhallen, R. J., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Bargary, G., & Mollon, J. D. (2017). General and specific factors in the processing of faces. *Vision Research, 141,* 217–227. https://doi.org/10.1016/j.visres.2016.12.014

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PloS One, 9*(8), e103510. https://doi.org/10.1371/journal.pone.0103510

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

World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects (2013). *JAMA, 310*(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053

5 Studie 3 - Individual differences in face recognition: Modulating effects of attention, processing speed, working memory, and self-reported face reocgnition ability

Petersen, L. A. (in preparation). Individual differences in face recognition: Modulating effects of attention, processing speed, working memory, and self-reported face recognition ability.

In diesem Kapitel wird Studie 3 vorgestellt, die unter dem Arbeitstitel „*Individual differences in face recognition: Modulating effects of attention, processing speed, working memory, and self-report face recognition ability*“ in einem Manuskript vorbereitet wurde, um es nach einer Datennacherhebung bei einer Fachzeitschrift einzureichen. Nachfolgend werden in Abschnitt 5.1 der theoretische Hintergrund, die Methodik und die Ergebnisse zusammenfassend dargestellt sowie einige Diskussionspunkte angesprochen. Das vorbereitete, englischsprachige Manuskript zur Studie 3 ist in Abschnitt 5.2 zu finden, sodass dort detaillierte Informationen zu den hier zusammengefassten Inhalten nachgelesen werden können.

5.1 Zusammenfassung Studie 3

Theoretischer Hintergrund

Die Gesichtsverarbeitungs- und Gesichtserkennungsfähigkeiten sind sehr heterogen von unterdurchschnittlich bis überdurchschnittlich verteilt (Bate et al., 2018; Fysh et al., 2020; McCaffery et al., 2018). Auch innerhalb der Gruppe von Super-Recognizern gibt es intraindividuelle Unterschiede (Bate et al., 2018; Belanova et al., 2021; Ramon et al., 2019). So zeigen nicht alle Super-Recognizer mit einem überdurchschnittlichem Personengedächtnis (zum Beispiel im CFMT+; Russell et al., 2009) auch überdurchschnittliche Personenvergleichsfähigkeiten (zum Beispiel im GFMT-S; Burton et al., 2010). Die kognitiven Einflussfaktoren auf die Gesichtsverarbeitungs- und Gesichtserkennungsfähigkeiten sind noch nicht vollständig aufgedeckt. Da die Gesichtsverarbeitung als kognitiver Prozess

(Bruce & Young, 1986) mit anderen kognitiven Prozessen verbunden sein sollte (Gignac, Shankaralingam, Walker, & Kilpatrick, 2016; Wilhelm et al., 2010; Wilmer, 2017; Wilmer et al., 2014), wurde bereits der Zusammenhang zur allgemeinen Intelligenz untersucht. Niedrige Korrelationen zwischen Leistungen in Gesichtserkennungstests (zum Beispiel CFMT; Duchaine & Nakayama, 2006) und Leistungen in verschiedenen Intelligenztests weisen darauf hin, dass die Gesichtsverarbeitung ein kognitiv spezifischer und von der allgemeinen Intelligenz unabhängiger Prozess ist (unter anderem Palermo, O'Connor, Davis, Irons, & McKone, 2013; Shakeshaft & Plomin, 2015). Es wäre aber möglich, dass spezifischere kognitive Subprozesse in Verbindung zur Gesichtsverarbeitung stehen.

Aus der Literatur lassen sich die Aufmerksamkeit (Bobak et al., 2017; Bruce & Young, 1986; Palermo & Rhodes, 2007; Wilhelm, Hildebrandt, & Oberauer, 2013), die Verarbeitungsgeschwindigkeit (Megreya & Burton, 2006; Wilhelm et al., 2010) und das Arbeitsgedächtnis (Gambarota & Sessa, 2019; Wilhelm et al., 2010) als mögliche Einflussfaktoren ableiten. So zeigen Maße kognitiver Subprozesse zwar eher niedrige Korrelationen zur Gesichtserkennungsleistung in Tests (zum Beispiel CFMT in Gignac et al., 2016; GFMT-S in McCaffery et al., 2018), doch könnten auch niedrige Korrelationen in der Summe zur Varianzaufklärung von Testleistungen beitragen und Hinweise zum Zusammenspiel beteiligter kognitiver Prozesse liefern. Im Kontext der Überprüfung der psychometrischen Qualität des CFMT+ und des GFMT-S lassen sich niedrige Korrelationen der Testleistungen zu Maßen kognitiver Prozesse außerdem als divergentes Validitätsargument interpretieren, da die Testkonstrukte zwar im Sinne kognitiver Fähigkeiten verwandt, jedoch nicht identisch sind (Campbell & Fiske, 1959; Kline, 2005). Neben kognitiven Subprozessen können auch andere Konstrukte oder die Messmethode die Gesichtserkennungsleistung modulieren. In der Literatur wird die Messung von Gesichtserkennungsfähigkeiten über die Selbsteinschätzung in Einzelitems oder mehrdimensionalen Fragebögen diskutiert

(Bobak et al., 2019; Bobak, Pampoulov et al., 2016; Shah, Gaule, Sowden, Bird, & Cook, 2015). Ein mehrdimensionaler Fragebogen („*Stirling Face Recognition Scale*“ = SFRS) von Bobak et al. (2019) zeigte bereits eine signifikante mittlere Korrelation zum Testergebnis im CFMT+ ($r(96) = .36$, $p < .01$). Studie 3 hat daher zusammenfassend die folgenden zwei Forschungsfragen untersucht:

1. In welchem Zusammenhang steht die Testleistung im CFMT+ zu den kognitiven Subprozessen Aufmerksamkeit, Verarbeitungsgeschwindigkeit und Arbeitsgedächtnis sowie zu der selbst eingeschätzten Gesichtserkennungsfähigkeit in einem Fragebogen?
2. In welchem Zusammenhang steht die Testleistung im GFMT-S zu den kognitiven Subprozessen Aufmerksamkeit, Verarbeitungsgeschwindigkeit und Arbeitsgedächtnis sowie zu der selbst eingeschätzten Gesichtserkennungsfähigkeit in einem Fragebogen?

Methodik

In zwei Teilen absolvierten 55 Proband:innen (63.63% Frauen, $M_{\text{Alter}} = 27.95$ Jahre, 19 bis 56 Jahre alt) die Tests dieser Studie (Erhebungszeitraum: 2018 bis 2019). Im ersten Studienteil wurden der CFMT+ und der GFMT-S online durchgeführt (wie in Studie 1, siehe Abschnitt 3.1). Beim zweiten Studienteil bearbeiteten die Proband:innen in einem Raum am Institut für Psychologie der Christian-Albrechts-Universität zu Kiel angeleitet durch eine Testleiterin folgende papierbasierten Aufgaben in fester Reihenfolge: den d2-R zur Messung der Aufmerksamkeit (Brickenkamp, Schmidt-Atzert, & Liepmann, 2010), die Untertests zum numerischen und verbalen Arbeitsgedächtnis sowie zur Verarbeitungsgeschwindigkeit aus dem WAIS-IV (Wechsler, 2012), und den ins Deutsche übersetzte „*Stirling-Face-Recognition-Scale*“-Fragebogen (selbst eingeschätzte Gesichtserkennungsfähigkeit; Bobak et al., 2019). Am Ende des zweiten Studienteils wurden die Proband:innen gebeten, noch einige Kontrollfragen für die Auswertung der kognitiven Tests und zu möglichen Störungen während der Aufgabenbearbeitung zu beantworten. Für die statistischen Analysen wurden die Rohwerte des

CFMT+ und des GFMT-S in T-Werte gemäß der Normtabelle für die Onlinetestdurchführung aus Petersen und Leue (2021; *Supplement* S8 und S11) umgerechnet. Für die kognitiven Aufgaben (d2-R, Untertests WAIS-IV) wurden die altersspezifischen Normwerte aus den jeweiligen Manualen verwendet (Standardwerte für den d2-R, Wertpunkte für die Untertests im WAIS-IV sowie IQ-Werte für den Gesamtwert der Skala zum Arbeitsgedächtnis und der Skala zur Verarbeitungsgeschwindigkeit). Da es für den SFRS keine Normwerte gibt, wurden die Summenwerte für den Gesamtwert und die Skalenwerte *z*-standardisiert.

Ergebnisse

Da für keine der Testwerte eine Normalverteilung angenommen werden konnte, wurden Spearman-Rank-Korrelationen berechnet. Es gab keine signifikanten Korrelationen zwischen den Testleistungen im CFMT+ (T-Werte) und den Leistungen in den kognitiven Aufgaben (verschiedene Normwerte). Ebenso gab es keine signifikanten Korrelationen zwischen den Testleistungen im GFMT-S (T-Werte) und den Leistungen in den kognitiven Aufgaben (verschiedene Normwerte). Einige Korrelationen zeigten aber die erwartete positive Richtung (siehe *Results* in Abschnitt 5.2). So zeigten sich die höchsten Korrelationen zwischen den Leistungen im CFMT+ und dem verbalen Arbeitsgedächtnis (*Rho*(55) = .19) sowie den Leistungen im GFMT-S und dem verbalen Arbeitsgedächtnis (*Rho*(55) = .21). Des Weiteren zeigten die T-Werte des CFMT+ eine Korrelation von .15 zum Genauigkeitsmaß F% im d2-R und die T-Werte des GFMT-S eine Korrelation von .11 zum Arbeitsgedächtnis-Gesamtwert. Im Gegensatz zur Erwartung wurden negative Korrelationen zwischen den Leistungen in den Gesichtserkennungstests und den zwei Untertests zur Verarbeitungsgeschwindigkeit gefunden (*Rho*(55) = -.13 bis -.08). Wie erwartet korrelierten die Leistungen im CFMT+ (T-Werte) und im GFMT-S (T-Wert) signifikant positiv mit dem *z*-standardisierten Gesamtwert des Fragebogens SFRS (*Rho*(55) = .47 und .49). Mit Cronbach’s Alpha von .90 und McDonald's Omega von .91 ist die Reliabilität des übersetzen Fragebogens als sehr gut zu bewerten

(vgl. Bühner, 2011). Zusätzlich wurde noch eine Hauptkomponentenanalyse (PCA) nach Prüfung der statistischen Voraussetzungen für die 20 Items des SFRS berechnet (siehe *Results* in Abschnitt 5.2). Die PCA favorisierte fünf Komponenten und konnte daher nicht die Zwei-Komponenten-Struktur der englischsprachigen Fragebogenversion von Bobak et al. (2019) replizieren. Abschließend wurden mit allen Variablen (*z*-standardisiert) lineare Regressionsmodelle berechnet (*Enter*-*Stepwise*-Methode), um die Gesamtvarianzaufklärung am Testergebnis im CFMT+ und im GFMT-S besser beurteilen zu können. So zeigte sich, dass der GFMT-S als Prädiktor für das Testergebnis im CFMT+ genau so viel Varianz aufklärt wie alle kognitiven Aufgaben mit der Selbsteinschätzungsfähigkeit zusammen. Im Gesamtmodell aller Variablen für das Kriterium CFMT+ lag das *adjusted* R^2 bei .36. Im Gesamtmodell aller Variablen für das Kriterium GFMT-S lag das *adjusted* R^2 bei .35.

Diskussion

Insgesamt konnten die Erwartungen für den Zusammenhang zwischen den Testwerten in den Gesichtserkennungstests und den Testwerten der anderen kognitiven Aufgaben nur für das Arbeitsgedächtnis, insbesondere für das verbale Arbeitsgedächtnis erfüllt werden (Forschungsfragen 1 und 2). Daher scheint es sinnvoll, das Arbeitsgedächtnis weiterhin in separate Funktionsbereiche nach zum Beispiel Oberauer et al. (2003) im Kontext der Gesichtserkennungsfähigkeit zu untersuchen. Allerdings lassen sich die Ergebnisse aufgrund der kleinen Stichprobengröße in dieser Studie nur eingeschränkt interpretieren und keine der Korrelationen wurde signifikant. Erst ab einer Stichprobengröße > 101 würde eine Korrelation von .21 (Korrelation zwischen den Leistungen im GFMT-S und im Untertest zum verbalen Arbeitsgedächtnis) mit einer Power von .80 signifikant werden (G*Power; Faul, Erdfelder, Buchner, & Lang, 2009). Interpretiert man jedoch vorsichtig die Richtung der Korrelationen der d2-R-Maße, könnte man sagen, dass Personen mit höheren Werten in den Gesichtserkennungstests sorgfältiger arbeiten (F%), aber auch langsamer sind (BZO) und eine

geringere Konzentrationsleistung (KL) haben. Das langsamere Arbeiten im d2-R (BZO) passt zu den negativen Korrelationen zwischen den Leistungen im CFMT+ und den Leistungen in den WAIS-IV Untertests zur Verarbeitungsgeschwindigkeit. Eine mögliche Erklärung könnte sein, dass Menschen mit hohen Werten in der Gesichtserkennung nicht langsamer arbeiten, sondern nur mehr Zeit für ihre Entscheidungen brauchen. Zukünftige Forschung könnte daher die Reaktionszeiten im CFMT+ sowie im GFMT-S im Zusammenhang mit der Verarbeitungsgeschwindigkeit untersuchen.

Ergänzend können alle Korrelationen zwischen den Leistungen im CFMT+/GFMT-S und den kognitiven Aufgaben als divergente Validitätsargumente für den CFMT+ und den GFMT-S interpretiert werden. Die Gesichtserkennungstests und die kognitiven Aufgaben erfassen inhaltlich unterschiedliche kognitive Konstrukte (vgl. Campbell & Fiske, 1959) und deren Korrelationen unter .30 können daher als divergente Validität ausgelegt werden (Kline, 2005). Des Weiteren konnte in dieser Studie in Übereinstimmung mit Bobak et al. (2019) der positive Zusammenhang zwischen der Gesichtserkennungsleistung im CFMT+ und der selbst eingeschätzten Gesichtserkennungsfähigkeit in einem Fragebogen (SFRS) repliziert werden, was als konvergentes Validitätsargument interpretiert werden kann (vgl. Campbell & Fiske, 1959; Kline, 2005). Insgesamt sprechen die Varianzerklärungen von 36% (CFMT+ als Kriterium) bzw. 35% (GFMT-S als Kriterium) dafür, dass es noch mehr Einflussfaktoren geben muss, die die individuellen Gesichtserkennungsleistungen erklären. Des Weiteren wird eine Datennacherhebung angestrebt, um die Mindeststichprobengröße für eine Hauptkomponentenanalyse bzw. lineare Regressionen mit bis zu 7 Prädiktoren zu erreichen (> 100; Field, 2011) sowie um die Analysen mit einer größeren Power durchführen zu können. Eine a-priori G*Power-Analyse für eine lineare multiple Regression mit sieben Prädiktoren ergibt $N = 115$ bei einer mittleren Effektstärke und einer Power von .85 (Faul et al., 2009).

5.2 Englischsprachiges Manuskript der Studie 3

Es folgt das vorbereitete Manuskript zur dritten Studie unter dem Arbeitstitel „*Individual differences in face recognition: Modulating effects of attention, processing speed, working memory, and self-reported face recognition ability*", welches nach Datennacherhebung bei einer Fachzeitschrift eingereicht werden soll. Das englischsprachige Manuskript enthält ein eigenes Literaturverzeichnis am Ende des Abschnitts.

Petersen, L. A. *(in preparation*). Individual differences in face recognition: Modulating effects of attention, processing speed, working memory, and self-reported face recognition ability.

Unpublished manuscript. Institute of Psychology, Kiel University

Individual Differences in Face Recognition: Modulating Effects of Attention, Processing Speed, Working Memory, and Self-Reported Face Recognition Ability

Lara Aylin Petersen

Kiel University

Abstract

The performance of individuals in face processing tasks (e.g., face recognition tests) is very heterogeneous. Face processing is a specific cognitive ability and the involved processes of the nomological network have not yet been elucidated. We investigated attention, processing speed, working memory, and self-reported face recognition ability as modulating factors that can be derived from cognitive research. In our study, we used the CFMT+ (face memory) and the GFMT-S (face matching) to measure face recognition ability (N = 55). The CFMT+ and the GFMT-S results showed non-significant small negative correlations with attention accuracy scores and processing speed. Moreover, non-significant small positive correlations were observed with verbal working memory and significant positive medium correlations with self-reported face recognition ability. After discussing the results in the context of previous cognitive research, we interpret preliminary convergent and divergent validity evidence for the CFMT+ and the GFMT-S. Finally, we provide ideas for future research.

Words 150

Keywords: face recognition, attention, processing speed, working memory, self-report

Introduction

The performance of individuals in face processing tasks (e.g., face recognition tests) is very heterogeneous (e.g., Bate et al., 2018; Fysh, Stacchi, & Ramon, 2020; McCaffery, Robertson, Young, & Burton, 2018). Individual differences also appear to modulate face processing performance within superior performance ("Super-Recognizers" = SR; Bate et al., 2018; Ramon, Bobak, & White, 2019; Russell, Duchaine, & Nakayama, 2009). However, SRs perform better than controls in face memory and face matching tasks (e.g., Bate et al., 2018; Petersen & Leue, 2021). The classification of SRs is predominantly determined by their short-term face memory ability (e.g., with the Cambridge Face Memory Test Long, CFMT+; Bobak, Pampoulov, & Bate, 2016; Petersen & Leue, 2021; Ramon et al., 2019; Russell et al., 2009) and their face matching ability (e.g., with the Glasgow Face Matching Test Short, GFMT-S; Burton, White, & McNeill, 2010; Davis, Bretfelean, Belanova, & Thompson, 2020; Petersen & Leue, 2021; Verhallen et al., 2017). The modulating factors that explain the differences in face processing performance remain undetermined. With face processing being related to other cognitive processes (Gignac, Shankaralingam, Walker, & Kilpatrick, 2016; Wilmer, 2017; Wilmer, Germine, & Nakayama, 2014), researchers have already investigated the cognitive specificity of face processing and general intelligence (e.g., fluid or nonverbal). The face processing ability is considered not to reflect a general cognitive functioning. This assumption is supported by small positive or non-significant correlations between face processing tasks and general cognitive measurements (range: -.08 to .21) in samples of N = 40 to 718 (Davis et al., 2011; Palermo, O'Connor, Davis, Irons, & McKone, 2013; Peterson & Miller, 2012; Shakeshaft & Plomin, 2015; Wilmer et al., 2014). Therefore, Wilmer (2017), who summarized the research on individual differences in face recognition, recommended examining more specific cognitive subprocesses to explain how individual differences of face recognition arise. In this line, we aimed to investigate cognitive subprocesses derived from cognitive models and research (Bruce

& Young, 1986; Gignac et al., 2016; Megreya & Burton, 2006; Palermo & Rhodes, 2007; Wilhelm et al., 2010). This aforementioned literature shows that executive functions such as attention, processing speed, and working memory are involved in perceiving, learning, and recognizing faces. Therefore, different cognitive subprocesses are likely to account for performance differences in face recognition tasks (e.g., the CFMT+ and the GFMT-S). However, in addition to cognitive factors, other factors could also modulate face recognition performance (e.g., Big Five personality traits; Megreya & Bindemann, 2013; Satchell, Davis, Julle-Danière, Tupper, & Marshman, 2019). Based on self-reports of SRs about their abilities (e.g., Russell et al., 2009), it seems possible that SRs have extraordinary self-evaluation abilities. A rather high correlation ($r(96) = .36$) was reported between the CFMT+ and a self-reported face recognition ability questionnaire (Stirling Face Recognition Scale = SFRS; Bobak, Mileva, and Hancock, 2019). A reliable questionnaire in several languages might be useful as a screening tool for research and in police applications (e.g., personnel recruitment of police officers). Therefore, the aim of our study was to investigate whether attention, processing speed, working memory, and the self-reported face recognition ability correlate to the performance in face recognition tests (i.e., the CFMT+ and the GFMT-S).

Face processing, face recognition, and cognitive subprocesses

Face processing ability is especially important when faces should be identified, e.g., in real life settings or in face recognition tests (McCaffery et al., 2018). Face identification is an important part of face memory tasks. A face should be memorized and later compared with other faces. In addition, face identification is an important part of face matching tasks. A face should be compared to another simultaneously presented face and checked for a match. Researchers investigated the relationship of face memory and face matching tasks to explore the underlying cognitive structure of face processing abilities. Due to rather high correlations between face memory and face matching tasks, some researchers assumed a common face

factor "f" underlying the face processing ability (about r = .45 in N = 28 – 397: Fysh, 2018; McCaffery et al., 2018; Verhallen et al., 2017; Wilhelm et al., 2010). In conjunction with small correlations to general intelligence (see Table 1 for some examples of general intelligence results measured by Catell's Culture Fair Intelligence Test (Catell, 1963), face processing has been considered to have its own cognitive structure (e.g., Bruce & Young, 1986; Wilhelm et al., 2010) and its own neural system (e.g., Haxby, Hoffman, & Gobbini, 2000). However, a variance elucidation of 23% (Verhallen et al. 2017) indicates that beyond factor "f" other factors are modulating the performance as well in face processing tasks.

The importance of investigating other modulating factors of face recognition in SRs increased during recent years (Bate et al., 2018; Fysh et al., 2020; Ramon et al., 2019; Ramon, 2021; Wilmer, 2017). SRs have been considered to have superior face processing abilities and they are usually classified by their performance in face memory tests, particularly the CFMT+ (Bobak et al., 2016; Davis et al., 2020; Ramon et al., 2019; Russell et al., 2009). However, a definition of superior face recognition abilities or a standard diagnostic tool of SR assessment has not yet been finalized (Bate, Portch, & Mestry, 2021; Ramon et al., 2019; Ramon, 2021). Not only face memory but also face matching has been associated with superior face recognition ability (e.g., Ramon et al., 2019). In the SR context, the GFMT-S (Burton et al., 2010) is used to measure face matching abilities (e.g., Belanova, Davis, & Thompson, 2021; Ramon et al., 2019). However, different performances in different face recognition tasks were observed within controls and within the group of superior performer (Bate et al., 2018; Belanova et al., 2021; Fysh et al., 2020; Ramon et al., 2019). Therefore, by investigating modulating factors, the research aims to find out how the performance differences in face recognition may occur.

Table 1

Overview of some studies that reported correlations between face recognition tests (CFMT, GFMT-S, other face matching tasks) and cognitive tasks.

Face matching or memory task	Cognitive test or task	Correlations with *N* and *CI*s (when available)	Reference for reported correlation
CFMT (Duchaine & Russell, 2006), face memory	Verbal Paired-Associates Memory Test, VPAM (Verbal memory)	*r*(50) = .12 [-.16, .38][1]	Wilmer et al., 2012
	Abstract Art Memory Test, AAM (Visual memory)	*r*(20) = .26 [.21, .31][a]	Wilmer et al., 2012
	Catell's Culture Fair Intelligence Test (Scale 3)	*r*(63) = -.08	Davis et al., 2011
	Catell's Culture Fair Intelligence Test – series (Scale 3)	*r*(211) = .03 – .14*	Gignac et al., 2016
	Catell's Culture Fair Intelligence Test – matrices (Scale 3)	*r*(211)= .16* – .19*	Gignac et al., 2016
	Digit Span backward (Working memory)	*r*(211) = .03 – .07	Gignac et al., 2016
	Word span backward (Working memory)	*r*(211) = .13* – .23*	Gignac et al., 2016
	Wechsler Abbreviated Scale of Intelligence (WASI), Matrix Reasoning	*r*(45) = .21	Peterson & Miller (2012)
	Navon local task (Visual object processing)	*r*(105) = .03 [-.16; .22]	McCaffery et al., 2018
	BADS Card Sorting (Executive function, cognitive flexibility)	*r(*105) = .12 [-.08; .30]	McCaffery et al., 2018
GFMT-S (Burton et al., 2010), face matching	Navon local task (Visual object processing)	*r*(105) = .27** [.01; .44]	McCaffery et al., 2018
	BADS Card Sorting (Executive function, cognitive flexibility)	*r*(105) = .20*	McCaffery et al., 2018
Other: Matching upright unfamiliar faces	Perceptual speed: Finding A's/ number comparisons/ identical pictures	*r*(30) = .46*/ .36 / .65**	Megreya & Burton, 2006

Note. * $p < .05$; ** $p < .01$; CFMT = Cambridge Face Memory Test (Duchaine & Nakayama, 2006); GFMT-S = Glasgow Face Matching Test Short (Burton et al., 2010); CI = confidence interval. [a] we did not find a *p*-value for this correlation.

In the literature, mostly small correlations between face memory or face matching tasks and other cognitive tasks can be found (see Table 1 for examples of the CFMT; Duchaine & Nakayama, 2006). So far, research has almost not investigated the relation between face recognition tasks and cognitive subprocesses. Consequently, we could not find correlations between the performances in the CFMT+ and in cognitive subprocesses tasks in the literature, except for the easier version CFMT. Likewise, we could only find a few results for the GFMT-S. Therefore, more data is needed on the CFMT+ and the GFMT-S.

Attention (i.e., concentration) is a directional component to guide active recognition (Bruce & Young, 1986; Palermo & Rhodes, 2007; Wilhelm, Hildebrandt, & Oberauer, 2013). In their eye tracking study, Bobak, Parris, Gregory, Bennetts, and Bate (2017) discussed a more efficient spread of spatial attention across faces for SRs. Attention is also closely related to processing speed (i.e., perceptual or mental speed) and working memory (Bühner, Krumm, Ziegler, & Pluecken, 2006; Wilhelm et al., 2013).

Megreya and Burton (2006) reported significant medium to large correlations between an unfamiliar face matching task and perceptual speed (see Table 1). This report corresponds to the results of Wilhelm et al. (2010), who published one of the most comprehensive studies examining the relation of face cognition variables (face perception, face memory, and face cognition speed) and cognitive processes (mental speed, object cognition, memory, and general cognitive ability). Wilhelm et al. (2010) found the strongest connection between mental speed (processing speed) and face cognition speed.

Furthermore, working memory as the central executive of cognitive processes (Logie, Camos, & Cowan, 2020; Oberauer, Süß, Wilhelm, & Wittman, 2003) was one of the general cognitive variables in the study of Wilhelm et al. (2010). According to the facet model of Oberauer et al. (2003), working memory can be divided into three content categories: visual-spatial, numerical, and verbal material. The numerical and verbal working memory has received

little attention in the research of face processing performance (for a review of encoding faces in visual working memory see Gambarota & Sessa, 2019). Previous research reports one non-significant small correlation between the CFMT (Duchaine & Nakayama, 2006) and a digital span task measuring numerical working memory ($r(211) = .03 – .07$; Gignac et al., 2016).

In summary, we assumed to find significant small positive correlations between the CFMT+ performance (face memory) and attention (Hypothesis 1a), processing speed (Hypothesis 1b), and working memory (Hypothesis 1c). Similarly, we assumed to find significant small positive correlations between the GFMT-S performance (face matching) and the above mentioned cognitive subprocesses (Hypothesis 2a–c).

Self-reported face recognition ability

However, besides cognitive subprocesses, other constructs or measurement factors can be discussed in the context of differences in face processing ability (e.g., Big Five personality traits: Megreya & Bindemann, 2013; Satchell et al., 2019). Based on self-reports about their cognitive abilities (e.g., Russell et al., 2009), SRs may also have extraordinary self-report face recognition abilities. Therefore, measuring the face recognition ability with self-report questionnaires or single-items is discussed in face recognition research (Bobak et al., 2016; Shah, Gaule, Sowden, Bird, & Cook, 2015; Verhallen et al., 2017). Although individuals normally show rather moderate self-report abilities using single-items (Bobak et al., 2016; Verhallen et al., 2017), a multidimensional questionnaire could more precisely measure self-reported face recognition abilities. Therefore, Bobak et al. (2019) developed the 20-items Stirling Face Recognition Scale (SFRS; with scale face memory and scale face perception) and reported a significant positive correlation between the SFRS-total raw score and the CFMT+ sum score ($r(96) = .36$, $p < .01$). Their results fit to Palermo et al. (2017), who developed a 77-items face processing questionnaire and reported a similar correlation to the CFMT sum score (Duchaine & Nakayama, 2006). Such a multidimensional questionnaire would be a useful

complement to the other tests (e.g., CFMT+) in SR research. Therefore, we aimed to replicate the findings of Bobak et al. (2019) with a German translation of the SFRS. We assumed to find a medium positive correlation between the CFMT+ performance and the German SFRS total score (Hypothesis 3). As the CFMT+ and the GFMT-S correlated positively (Petersen & Leue, 2021) and the SFRS had a subscale entitled face perception, we assumed to find a medium positive correlation between the GFMT-S performance and the German SFRS total score (Hypothesis 4).

Method

Participants

A total of N = 55 participants (35 females, 63.64%) with a mean age of 27.95 years (SD = 10.11, *range* 19 – 56 years) took part in this study. The sample consisted of 31 psychology students, eight students of other disciplines, and 16 employees with a higher educational level. The participants were recruited at the Kiel University or via mail invitation. All participants performed the face recognition tests for the first time and were not familiar with the cognitive tasks (except four participants who had performed the attention task shortly before in another study and therefore their attention scores were set missing; see Results). All participants were white Caucasian, reported normal or corrected to normal vision, and did not reported any neurological diseases. Psychology students completed the study for credit points and the other participants received a small monetary payment.

Materials

Face recognition

Cambridge face memory test long (CFMT+; Russell et al., 2009)

The CFMT+ is a standardized test for short-term face memory and was conducted online. The test asks participants to learn six male target faces (gray-scaled). Afterwards, they have to recognize the target faces (retrieved from memory) in forced-choice items. One face

out of three is known and participants have to choose the known face. For each correct decision a participant gets one point (102 items in total = maximal 102 points). The difficulty of the presented items increases over four sections. The CFMT+ does not have a time limit and can be usually completed in 20 minutes. Some test properties (e.g., Cronbach's Alpha = .92) for the Online CFMT+ are reported in Petersen and Leue (2021). Further description and examples of the items are reported in Russell et al. (2009).

Glasgow face matching test short (GFMT-S; Burton et al., 2010)

The short version of the GFMT (Burton et al., 2010) is a standardized face matching test and was conducted online. The GFMT-S asks participants to compare two faces in 40 items. Half face pairs have the same identity and half have a different identity. For each correct decision, the participant gets one point (maximal 40). The GFMT-S does not have a time limit and can be usually completed in 10 minutes. Some test properties (e.g., Cronbach's Alpha = .71) of the Online GFMT-S are reported in Petersen and Leue (2021). Further description and examples of the items are reported in Burton et al. (2010).

Stirling Face Recognition Scale (SFRS; Bobak et al., 2019) – German language

The scale comprises 20 face recognition-specific statements in relation to the experience of the participants' face recognition ability. Bobak et al. (2019) adapted the statements from the 20-item Prosopagnosia-Index (PI-20; Shah et al., 2015) and developed the SFRS with two scales: self-reported face memory and self-reported face perception. For example, the statement of Item 1 (scale face memory) reads: "I know exactly where I first met someone (work meetings, parties, etc.)". We translated the statements of the SFRS into the German language (backward forward) and presented them in a paper-pencil-based form. The response categories of the items range from strongly disagree (score = 1) to strongly agree (score = 5). The maximum possible score is 100 points. The SFRS includes half positive and half reversed scored items to refer to the low and high end of the face recognition spectrum. Answering the items have no time limit.

Participants are asked to answer the questions as honestly as possible. The reliability of the SFRS was good in Bobak et al. (2019) with Cronbach's Alpha = .88 and Split-Half Spearman-Brown corrected reliability = .89 (Bühner, 2011).

Cognitive subprocesses

We operationalized the cognitive subprocesses attention (A), processing speed (PS), and working memory (WM) as follows. Attention as an executive cognitive process (Bühner et al., 2006) was operationalized by norm scores of the d2-R *Test of Attention* (Brickenkamp, Schmidt-Atzert & Liepmann, 2010). We measured the verbal and numerical WM separate according to the facet model of Oberauer et al. (2003). We used the subtests for WM from the *Wechsler Adult Intelligence Scale – Forth Edition* (WAIS-IV; Wechsler, 2012). Further, we used the two subtests for PS from the WAIS-IV (Wechsler, 2012). Test properties are reported in Brickenkamp et al. (2010) for the d2-R and in Wechsler (2012) for the subtests of the WAIS-IV. All cognitive tests were presented paper-pencil-based and conducted by a trained investigator in a laboratory at Kiel University.

d2-R (Brickenkamp et al., 2010) [A]

The d2-R measures focused attention, i.e., concentration in tasks that require attention. The d2-R has a time limit and takes five minutes (including the instruction time). Participants are asked to cross out any letter "d" with two marks above or below (target object) in a row with letters p and d with too few or too many marks as distractors (14 rows in total). For each row participants have 20 seconds to cross out the target objects and to avoid the distractor objects. Three important scores can be calculated from the total crossed letters or errors in each row, KL: Attention performance (number of correctly crossed target objects minus number of errors), BZO: Performance speed (number of target objects completed in total), and F%: Relative error frequency, accuracy ((100 * number of errors)/ BZO). The reliability of the KL score and the BZO score were mainly excellent (Bühner, 2011) with Cronbach's Alpha

coefficients between .89 and .95 (Brickenkamp et al., 2010). The reliability of the F% score was good with Cronbach's Alpha coefficients between .80 and .91 (Brickenkamp et al., 2010).

WAIS-IV: Working Memory – Subtest Digit Span (Wechsler, 2012) [WM-DS]

This subtest measures the numerical WM and does not have a time limit. Participants are asked to listen verbally pronounced sequences of numbers by the experimenter. The test is divided into three parts with slightly different tasks. Participants are asked to repeat as many sequences of numbers as possible before the next part starts. In the first part (digital span forward), participants repeat the numbers in the same order as they have heard the numbers. In the second part (digital span backward), they repeat the numbers in the reversed order. In the third part (digital span sequential), they sort the heard numbers in an increasing order. The performances in all three parts yield a total score for the subtest.

WAIS-IV: Working Memory – Subtest Letter-Number-Sequencing (Wechsler, 2012) [WM-LNS]

This subtest measures the verbal WM (attention and mental control is needed to complete the task) and does not have a time limit. Participants are asked to listen to sequences of numbers and letters verbally presented by the investigator. Afterwards, they repeat the numbers in an ascending order and the letters in an alphabetical order.

The subtests WM-DS and WM-LNS yield in a total score for "Working Memory". The reliability for the "Working Memory" scale is reported with .94 (Cronbach's Alpha; see test manual; Wechsler, 2012) and can be rated as excellent (Bühner, 2011).

WAIS-IV: Processing Speed – Subtest Symbol Search (Wechsler, 2012) [PS-SS]

This subtest measures PS and have to be completed in two minutes. In rows of different symbols, participants say whether a target symbol appears or not. Each row has new target and distractor symbols.

WAIS-IV: Processing Speed – Subtest Coding (Wechsler, 2012) [PS-C]

This subtest measures PS and have to be completed in two minutes. Participants write down (draw) symbols under rows of numbers by using a given number-symbol code presented in the upper part of the task paper.

The subtests PS-SS and PS-C yield a total score for "Processing Speed". The reliability for the "Processing Speed" scale is reported with .88 (Test-Retest correlation, no Cronbach's Alpha is reported; see test manual; Wechsler, 2012) and can be rated as good (Gregory, 2014).

Procedure

The study consisted of two sections that were conducted on separate dates. With all tasks in this study requiring a cognitive load, we aimed at keeping motivation and acceptance for task performance. We also aimed at reducing confounding variables such as fatigue. For economic reasons, the face recognition tests were conducted online, since some participants had already completed the CFMT+ and the GFMT-S for another study (Petersen & Leue, 2021). Therefore, these participants were exclusively invited to conduct the cognitive tasks and the SFRS in our laboratory. Other participants (recruited at the university) first received the invitation to conduct the online tests. All participants conducted the first section of the study online via SoSci Survey (https://www.soscisurvey.de/en/index). At the beginning, all participants gave written informed consent and then answered questions on demographic data. Subsequently, the CFMT+ and the GFMT-S were performed in a fixed order, followed by control questions for the assessment of data quality. We ensured that the online data were of good quality by defining a-priori criteria, e.g., no participation by smartphone, no conduction late at night, no interruption longer than five minutes, and no technical problems. We only invited participants to the second section of the study who fulfilled the a-priori criteria of data quality. Both test results were presented at the end of the first study section. The first online section took about 45 minutes.

In the second section of the study, the participants performed the cognitive tasks and the SFRS in a fixed order in a laboratory at the university, guided by an investigator: d2-R, WM-DS, PS-SS, WM-LNS, PS-C, and SFRS. This fixed order was chosen according to the information in the test manuals (d2-R at the beginning, WAIS-IV subtests in the order they appear in the WAIS-IV) and the questionnaire with the lowest cognitive load last. In the end, the participants answered an additional questionnaire with control questions for the cognitive tasks, e.g., to ensure they did not suffer from neurological diseases, which could influence the results. The second part of the study took about 55 minutes.

Using the correct responses, the CFMT+ and the GFMT-S performances were scored dichotomously for each item and resulted in a total test score of maximum 102 points and maximum 40 points, respectively, for each participant. For the statistical analyses, we transformed the raw scores of the CFMT+ and the GFMT-S into T-values based on the norm tables for online test presentation published in Petersen and Leue (2021; Supplement S8 and S11). The cognitive tasks were scored in accordance with the test manuals using the age-specific norm tables in Brickenkamp et al. (2010; d2-R norm sample, N = 4024; Standard-values) and in Wechsler (2012; WAIS-IV subtests norm sample, N = 1650, Point-scale-values and IQ-values). No norm tables were available for the SFRS, so we *z*-standardized the raw values for the analyses. All analyses were conducted with IBM SPSS Statistics (Version 26) for Windows.

Results

As further data is to be collected, the following results should be regarded as preliminary. Particularly due to the small sample size and the associated limited power (see power analyses below), the results should be interpreted with the necessary care.

Descriptive statistics, H1a–c and H2a–c: Face recognition and cognitive subprocesses

Table 2 shows the descriptive statistics for the CFMT+ (T-values), the GFMT-S (T-values), and the cognitive tasks (different norm values based on the manuals). Normal distribution of all scores were tested with the Kolmogorov-Smirnov test because the sample size was larger than $N = 50$ (Bühner, 2011). All scores were not normally distributed (a-priori alpha level: $p = .20$). Therefore, we calculated Spearman's Rho correlations (Table 3) to examine the hypotheses. Hypotheses 1a–c (CFMT+) and 2a–c (GFMT-S) could not be confirmed due to non-significant correlations between the CFMT+/ GFMT-S T-values and the norm scores of the cognitive tasks. However, some correlations showed the predicted positive direction and are similarly large to the reported correlations in Table 1. The CFMT+ and the GFMT-S showed the highest correlations to WM-LNS (verbal working memory; CFMT+: *Rho* = .19; GFMT-S: *Rho* = .21). In line with our expectations, the CFMT+ showed a positive correlation to d2-R F% (accuracy; *Rho* = .15) and the GFMT-S showed a positive correlation to WM total score (*Rho* = .11), though not significant. Contrary to our expectation, some correlations were negative. Both processing speed tasks (PS-SS; PS-C) showed negative correlations to the performance in the CFMT+ and the GFMT-S (*Rho* = -.13 to -.08). However, the cognitive tasks showed predominantly positive significant correlations to each other. The correlation between the CFMT+ T-values and the GFMT-S T-values was large (*Rho* = .61; Cohen, 1988).

Table 2

Descriptive statistics for the face recognition tests (CFMT+ T-values, GFMT-S T-values) and the cognitive tasks (see Material and Procedure; Column order according to task order).

	CFMT+	GFMT-S	F%	BZO	KL	WM-DS	PS-SS	PS-CC	WS-LMS	WM total	PS total
N	55	55	51[a]	51[a]	51[a]	55	55	55	55	55	55
M	47.29	47.93	103.57	104.77	103.49	10.36	11.51	13.11	10.62	102.71	113.00
SD	10.26	9.10	10.07	11.07	10.99	2.30	3.11	2.54	2.73	12.47	13.95
Skewness	1.13	-.03	-0.70	-0.02	-0.51	-0.21	0.53	-0.03	0.33	-0.03	0.30
Kurtosis	1.96	-0.80	0.79	-0.37	0.86	-0.92	-0.09	-0.05	-0.57	-0.65	-0.23
KS	.11*	.12*	.16*	.09*	.07*	.20*	.15*	.13*	.15*	.08*	.12*
Minimum	31	30	75	79	70	6	6	7	6	77	83
Maximum	79	63	126	130	130	15	19	19	17	128	146

Note. CFMT+ = Cambridge Face Memory Test Long, T-values; GFMT-S = Glasgow Face Matching Test Short, T-values; KS = Kolmogorov-Smirnov test for normality; F%, BZO, KL = norm scores of the d2-R attention task (Standard-values; Brickenkamp et al., 2010); WM-DS, WM-LMS = Working memory tasks of the Wechsler Adult Intelligence Scale – Fourth edition (Point-scale-values; Wechsler, 2012); PS-SS, PS-CC = Processing speed tasks of the Wechsler Adult Intelligence Scale – Fourth edition (Point-scale-values; Wechsler, 2012); WM total = IQ norm score for the total subscale WM; PS total = IQ norm score for the total subscale PS; * $p < .20$ for KS test.

[a] = the results of four participants in the d2-R were set to missing as they performed the d2-R in another study within two weeks before our study conduction (risk of retest effects).

Table 3

Spearman's Rho correlation matrix for the face recognition tests (CFMT+ T-values, GFMT-S T-values) and the cognitive tasks (attention, processing speed, working memory with norm scores).

Tasks	1	2	3	4	5	6	7	8	9	10
1 CFMT+	—									
2 GFMT-S	**.61***	—								
3 d2R - F%[a]	.15	.10	—							
4 d2R - BZO[a]	-.08	-.15	**.43***	—						
5 d2R - KL[a]	-.11	-.09	**.64***	**.91***	—					
6 WM - DS	-.03	.01	**.46***	**.39***	**.48***	—				
7 PS - SS	-.13	-.08	.20	**.71***	**.65***	.23	—			
8 PS - C	-.09	-.10	**.38***	**.55***	**.62***	**.44***	**.54***	—		
9 WM - LNS	.19	.21	**.44***	.10	.22	**.50***	.02	.15	—	
10 WM total	.09	.11	**.53***	.27	**.40***	**.86***	.14	**.34***	**.86***	—
11 PS total	-.13	-.11	**.32***	**.72***	**.72***	**.40***	**.87***	**.87***	.09	**.27***

Note. $N = 55$; * $p < .05$, two-tailed; CFMT+ = Cambridge Face Memory Test Long, T-values; GFMT-S = Glasgow Face Matching Test Short, T-values; F%, BZO, KL = norm scores of the d2-R (attention task; Brickenkamp et al. 2010); WM-DS, WM-LMS = Working memory tasks of the Wechsler Adult Intelligence Scale – Fourth edition (Wechsler, 2012); PS-SS, PS-CC = Processing speed tasks of the Wechsler Adult Intelligence Scale – Fourth edition (Wechsler, 2012); WM total = IQ norm score for the total subscale WM; PS total = IQ norm score for the total subscale PS.

[a] = the results of four participants in the d2-R were set to missing as they performed the d2-R in another study within two weeks before our study conduction (retest effects, see manual of d2-R), therefore $N = 51$.

Descriptive statistics, H3 and H4: Face recognition and self-report ability

The descriptive statistics for the SFRS total score (*z*-standardized), the subscales memory (*z*-standardized), and the subscale perception (*z*-standardized) are shown in Table 4. Three participants showed a CFMT+ T-value above 70 (far above average; Petersen & Leue, 2021). These three participants also reached a maximum score of > 63 T-values in the GFMT-S (Petersen & Leue, 2021) and had *z*-standardized SFRS-total scores between .30 and 2.14, which shows quite a large range. Kolmogorov-Smirnov tests showed that all SFRS *z*-scores were not normally distributed (a-priori α-level: $p = .20$). Therefore, we calculated the Spearman's Rho correlations between the CFMT+ T-values and the *z*-standardized SFRS total score (*Rho*(55) = .47, $p < .001$), the *z*-standardized SFRS subscale memory scores (*Rho*(55) = .53, $p < .001$) and perception (*Rho*(55) = .30; $p < .05$). Hypothesis 3 was confirmed by the significant positive medium correlations ($.30 > r < .50$; Cohen, 1988). We further calculated the Spearman's Rho correlations between the GFMT-S T-values and the *z*-standardized SFRS total score (*Rho*(55) = .49, $p < .001$), the *z*-standardized SFRS subscale scores memory (*Rho*(55) = .46, $p < .001$), and perception (*Rho*(55) = .46; $p < .001$). Thereby giving us confirmation of Hypothesis 4.

The Spearman's Rho correlations between the SFRS *z*-scores were also significantly large (Cohen, 1988): SFRS total score to memory scale (*Rho*(55) = .91, $p < .001$); SFRS total score to perception scale (*Rho*(55) = .87, $p < .001$); memory scale to perception scale (*Rho*(55) = .64, $p < .001$). The Cronbach's Alpha reliability for the 20 SFRS items was .90 with a mean inter-item correlation of .32. The McDonald's Omega reliability (Hayes & Coutts, 2020) was .91.

Table 4

Descriptive Statistics for the SFRS total score (z-standardized) and the subscales memory/perception (z-standardized).

	Total score	Memory-scale	Perception-scale
M	.05	.08	.00
SD	1.00	.98	1.00
Skewness	-.26	-.07	-.59
Kurtosis	.41	.84	-.26
KS	.08*	.11*	.16*
Minimum	-2.54	-2.67	-2.31
Maximum	2.14	2.32	1.54

Note. $N = 55$; KS = Kolmogorov-Smirnov test of normal distribution; * $p < .20$ for KS test

Additional analyses: Principal Component Analysis for the SFRS

In addition, we performed a Principal Component Analysis (PCA) for the 20 SFRS items with Direct-Oblim-Rotation to examine the underlying structure of the SFRS. However, it must be taken into account that our sample size is below the recommended size for a PCA (i.e., 4 cases per item = 80 participants according to Little, 2013; $N > 100$ according to Field, 2011). The Kaiser-Meyer-Olkin measure of sampling adequacy was .82, representing a relatively good component analysis (Field, 2011). Bartlett's test of Sphericity was significant ($\chi^2 = 531.81$, $df = 190$, $p < .001$), indicating that correlations between items were sufficiently large for performing a PCA (Field, 2011). Examination of Kaiser's criteria for eigenvalues (eigenvalue > 1) and the Screeplot (see Figure 1) yielded empirical justification for five components. With eigenvalues > 2, a two-component solution is also a possible interpretation of the PCA results. We report the component loadings for the five-component solution and the original items in Table 5. The SFRS items translated into German with the instruction are shown

in the Supplement. The analyses of the item to component assignment in Table 5 did not replicate the two-component solution of Bobak et al. (2018). Moreover, we could not find a suitable explanation for the assignment of the items to the five components.

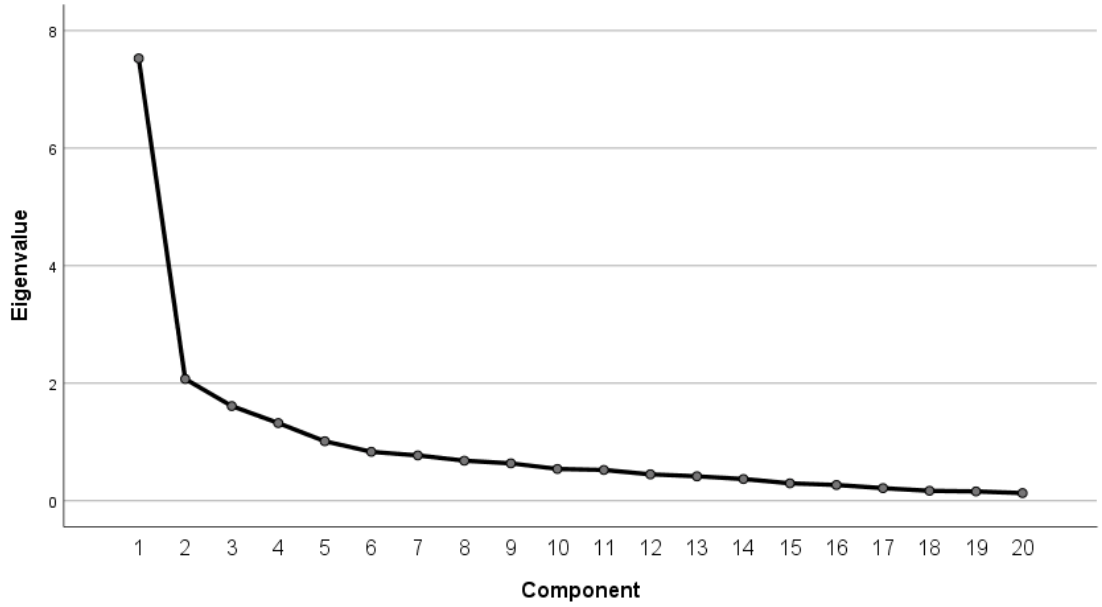


Figure 1. Screeplot of a Principal Component Analysis (PCA) with the 20 SFRS items translated into German.

Table 5

Rotated (oblim, direct) component loadings (PCA, eigenvalues > 1) for a five-component solution with Component 1 (C1, 37.63% of variance), Component 2 (C2, 10.36% of variance), Component 3 (C3, 8.05% of variance), Component 4 (C4, 6.61% of variance), and Complement 5 (C5, 5.06% of variance) supplemented by the original item formulations of Bobak et al. (2019).

Items	C1	C2	C3	C4	C5
Q1_M: I know exactly where I first met someone (work meetings, parties, etc.).	.31	**-.56**	-.37		
Q2_P: When I was at school I struggled to recognize my classmates.	**.52**			-.40	.43
Q3_M: I find it very easy to picture individual faces in my mind with great detail.	**.52**	-.60			
Q4_P: Anxiety about poor face recognition has led me to avoid certain social or professional situations.	.33	.43		**.61**	
Q5_M: I am better than most people at putting a "name to a face."	**.71**				
Q6_P: I often mistake people I have met before for strangers.	**.68**		-.32	-.35	
Q7_P: I tend to forget faces very quickly after seeing them.	**.83**				
Q8_M: I find it very easy to recognize familiar people when I meet them out of context (e.g., meeting a work colleague unexpectedly in the gym).	**.73**				-.37
Q9_M: Faces are enough for me to recognize people - I don't need to hear their voice or see their whole body.	**.67**		.32		
Q10_P: When people change their hairstyle, or wear hats, I have problems recognizing them.	**.59**		.34		
Q11_M: I am very good at spotting family resemblance in a group of unfamiliar people.	**.57**	-.49			
Q12_P: I sometimes have to warn new people I meet that I am "bad with faces."	**.63**	.55			

Table 5. (continuied).

Items	C1	C2	C3	C4	C5
Q13_M: I often recognize people I have met briefly before, but they have no idea who I am.	**.64**				
Q14_P: I have to try harder than other people to memorize faces.	**.86**				
Q15_M: I never forget a face.	**.72**		-.37		
Q16_P: I sometimes find movies hard to follow because of difficulties recognizing characters.		.39	-.40	**.62**	
Q17_M: I find it easy to recognize my family and friends from their childhood photographs.	.35		**.55**	.33	
Q18_P: I struggle to recognize people without hearing their voice.	**.59**	.31			
Q19_P: I often fail to recognize someone who knows me.	**.62**				
Q20_M: I easily recognize someone even if I can only see part of their face.	**.72**				

Note. N = 55; M = Items of the memory scale after Bobak et al. (2018); P = Items of the perception scale after Bobak et al. (2018); showing only salient component loadings > *.30.*

Additional analyses: Variance elucidation in linear regressions models

In addition to single correlations, the overall variance elucidation of all variables in this study is also of interest (cf. face factor "f"; Verhallen et al., 2017). Therefore, we calculated linear regression models with *z*-standardized variables using enter-stepwise-method to report the adjusted R^2s. Due to the small sample size, the following results should be interpreted with caution. A sample size of at least 80 participants is recommended with seven predictor variables in a linear regression for an effect of .20 and a power of .80 (G*Power; Faul, Erdfelder, Buchner, & Lang, 2009). The general recommendation with seven predictor variables in a linear regression provides a sample size of 106 (Field, 2011).

Model 1a was calculated using the CFMT+ (*z*-value) as a criterion and the following variables as predictors (all *z*-values): d2-R F%, d2-R BZO, d2-R KL, WM-total, PS-total,

SFRS total score. Model 1b added the GFMT-S (*z*-value) as additional predictor. Model 1a had an adjusted R^2 of .20 (R^2 = .21, $F(1, 49) = 13.08$, $p < .001$; SFRS total score, ß = .46, $p < .001$) and Model 1b with added GFMT-S had an adjusted R^2 of .36 (R^2 = .38, $F(2, 48) = 14.99$, $p < .001$; SFRS total score, ß = .26, $p < .05$; GFMT-S, ß = .46, $p < .001$). This shows that the GFMT-S contributes almost as much to the variance elucidation as the other tasks together. We calculated the same two linear regressions models for the GFMT-S (*z*-value) as criterion: Model 2a (cognitive tasks and SFRS) and Model 2b added CFMT+ as predictor variable. Model 2a had an adjusted R^2 of .18 (R^2 = .19, $F(1, 49) = 11.77$, $p < .001$; SFRS total score, ß = .44, $p < .001$) and Model 2b with CFMT+ had an adjusted R^2 of .35 (R^2 = .37, $F(2, 49) = 24.37$, $p < .001$; SFRS total score, ß = .22 , p =.09; CFMT+, ß = .47, $p < .01$). This shows that the CFMT+ contributes almost as much to the variance of the GFMT-S performance as the other tasks combined.

Discussion

In this study, four hypotheses concerning modulating factors on face recognition performance were investigated. Face recognition ability was measured by the CFMT+ (face memory test) and the GFMT-S (face matching test). The investigated modulating factors measured by different tasks were attention, processing speed, working memory, and the self-reported face recognition ability (measured by the SFRS questionnaire). The results showed non-significant small positive (verbal working memory and accuracy score of attention task) or negative correlations (both processing speed tasks, attention performance, attention performance speed) between the cognitive subprocesses tasks and the two face recognition tests. The Hypotheses 1a–c on CFMT+ and the Hypothesis 2a–c on GFMT-S could not be confirmed due to the non-significant results. As predicted, significant medium positive correlations between the CFMT+ T-values, the GFMT-S T-values, and the self-reported face

recognition ability (*z*-standardized SFRS scores) were found. Therefore, Hypotheses 3 and 4 were confirmed. We discuss the results and implications below.

The relation of cognitive subprocesses and face recognition performance

In accordance with previous studies on the CFMT (Duchaine & Nakayama, 2006) and different cognitive tasks (see Table 1 for examples), small negative or positive correlations were found for the CFMT+ and the cognitive tasks in this study. Contrary to our expectation, only the accuracy attention score (F%) of the well validated d2-R (Brickenkamp et al., 2010) showed a small positive correlation to the CFMT+ T-values and also to the GFMT-S T-values. The performance speed score (BZO) and the attention performance score (KL) of the d2-R showed small negative correlations. When carefully interpreting the direction of the non-significant correlations one could say individuals with higher scores in face recognition tests make relatively few errors (F%), but are also slower (BZO), and have a lower attention performance (KL). A possible modulating effect of processing information on face recognition performance fits to the negative correlations of the processing speed task in our study (*Rho*(55) = - .08 to -.13; see Table 3). However, a sample size of > 131 is required to reject the null hypotheses of no association with a correlation of -.13 to achieve a power of .80 (G*Power; Faul et al., 2009). If the direction of the negative correlations between face recognition and processing speed would be confirmed in a study with a larger sample size, this would be contrary to the results of Wilhelm et al. (2010). They found the strongest connection between face cognition speed (operationalized by a face short-term memory task and an old/new face identification task) and mental speed (processing speed). A possible explanation for our results is that individuals with higher face recognition performance are not slower in processing information but need more time for their decisions. We did not analyze the response times in the CFMT+ or the GFMT-S. Therefore, future research should investigate which explanation is more appropriate by using different processing speed tasks (to validate the results) and by analyzing the response times of

the CFMT+ or the GFMT-S items. Response times should preferably examined in a laboratory study, as there may be technical interferences online.

However, we investigated the working memory with a verbal and numerical working memory task. The correlations between the performances in the CFMT+ and the performances in the numerical (Table 3, WM-DS: *Rho*(55) = -.03) and the verbal (WM-LNS: *Rho*(55) = .19) working memory task fit to the results of Gignac et al. (2016) who reported similar significant correlations for the shorter CFMT. Therefore, it seems worthwhile to investigate working memory in its content categories according to the taxonomy of Oberauer et al. (2003) in context of face recognition performance differences. In addition, the measurements of the cognitive tasks should function as intended, since they showed high correlations among each other, which corresponds to the literature (Bühner et al., 2006; Wilhelm et al., 2010).

The relation of the self-report ability and face recognition performance

In accordance with the study of Bobak et al. (2019), we found a significant medium correlation (.30 to .50; Cohen, 1988) between the SFRS total *z*- score and the CFMT+ T-values (*Rho*(55) = .47;). The GFMT-S T-values also showed a significant medium correlation to the SFRS total *z*-score (*Rho*(55) = .49). Our translation of the questionnaire worked well due to the high Cronbach's Alpha (.90) and McDonald's Omega (.91) reliability coefficients. Therefore, we replicated the finding that multiple item scales can estimate the face recognition ability to some extent (Bobak et al., 2019; Palermo et al., 2017). However, we could not replicate the item to component assignment reported by Bobak et al. (2019; two-component solution for the SFRS) with our data. Even though the Kaiser-Meyer-Olkin criterion and Bartlett’s test of Sphericity suggested that the data were suitable for a PCA, our sample size is below the recommended size for a PCA (i.e., 4 cases per item = 80 participants according to Little, 2013; $N > 100$ according to Field, 2011). Therefore, the results of the PCA should be interpreted with caution.

However, as in the study of Bobak et al. (2019) our participants have briefly seen their face recognition test results online while completing the first section of the study. Therefore, they may have been not entirely unbiased in answering the questionnaire at the second part of the study. On the other hand, previous exposure to the contents of the questionnaire may also influence the face recognition test performance, e.g., test motivation or self-beliefs. It seems necessary to consider which direction of influence should be prevented. Future research could examine how large the effects of the test results knowledge or questionnaire exposure are. Furthermore, future research could examine how the face recognition ability (e.g., below average, average, far above average) is related to self-insights.

Predictors for face recognition test performance and test validity

Overall, the correlations and linear regression analyses with *z*-standardized variables showed that the self-reported face recognition ability and other face recognition tests are the strongest predictors for the test performance. However, all variables of this study together explained 36% of the variance (adjusted R^2, Model 1b) of the CFMT+ and 35% of the variance (adjusted R^2, Model 2b) of the GFMT-S. This finding implicates that more modulating factors exist. Future research should consider which factors might explain the performance differences in face recognition tests, such as the decision time for answering the face recognition test item or personality constructs (e.g., decision-making).

Finally, in the context of construct validity the correlations between the cognitive subprocesses tasks and the CFMT+ or the GFMT-S can be interpreted as preliminary divergent validity (American Educational Research Association, 2014; Campbell & Fiske, 1959). According to Kline (2005). A negative to small positive correlation (< .30; Cohen, 1988) can be rated as divergent (or discriminant) validity. In contrast, a medium correlation (between .30 and .50; Cohen, 1988) can be rated as sufficient for a convergent correlation. Thus, the correlations between the CFMT+ T-values /GFMT-S T-values to the SFRS total *z*-scores can

be interpreted as convergent validity ($r > .30$, Kline, 2005; same construct measured with different methods). Especially in an application where individual diagnostic decisions are made (such as the SR classification based on test results), it is important to use approved tests that fulfill standards in psychological assessment (American Educational Research Association, 2014) and to report validity findings.

Conclusion

In conclusion, the preliminary results of our study show that although cognitive subprocesses contribute only a small amount to the face recognition test performance, even these small modulating effects may provide hints about the cognitive relations. Furthermore, our results demonstrate that it might be worthwhile to assess face recognition ability not only with tests but also with multidimensional questionnaires. However, our results also show that there are more modulating factors regarding the face recognition test performance. Therefore, exploring individual differences in face recognition is worthwhile and future research should investigate more modulating factors.

References

American Educational Research Association; American Psychological Association; National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., . . . Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications, 3*, 22. https://doi.org/10.1186/s41235-018-0116-5

Bate, S., Portch, E., & Mestry, N. (2021). When two fields collide: Identifying "super-recognisers" for neuropsychological and forensic face recognition research. *Quarterly Journal of Experimental Psychology (2006)*, 17470218211027695. https://doi.org/10.1177/17470218211027695

Belanova, E., Davis, J. P., & Thompson, T. (2021). The part-whole effect in super-recognisers and typical-range-ability controls. *Vision Research*, *187*, 75–84. https://doi.org/10.1016/j.visres.2021.06.004

Bobak, A. K., Mileva, V. R., & Hancock, P. J. (2019). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology (2006)*, *72*(4), 872–881. https://doi.org/10.1177/1747021818776145

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of young British adults. *Frontiers in Psychology*, *7*, 1378. https://doi.org/10.3389/fpsyg.2016.01378

Bobak, A. K., Parris, B. A., Gregory, N. J., Bennetts, R. J., & Bate, S. (2017). Eye-movement strategies in developmental prosopagnosia and "super" face recognition. *Quarterly Journal of Experimental Psychology (2006)*, *70*(2), 201–217. https://doi.org/10.1080/17470218.2016.1161059

Brickenkamp, R., Schmidt-Atzert, L., & Liepmann, D. (2010). *d2-R–Aufmerksamkeits- und Konzentrationstest. (d2-R-Attention and concentration test)*. Göttingen: Hogrefe.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology (London, England: 1953), 77 (Pt 3)*, 305–327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte und erw. Aufl.). *PS Psychologie*. München: Pearson Studium. Retrieved from http://lib.myilibrary.com/detail.asp?id=404890

Bühner, M., Krumm, S., Ziegler, M., & Pluecken, T. (2006). Cognitive abilities and their Interplay. *Journal of Individual Differences*, *27*(2), 57–72. https://doi.org/10.1027/1614-0001.27.2.57

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*(1), 286–291. https://doi.org/10.3758/BRM.42.1.286

Catell, R. B. (1963). *The IPAT culture fair intelligence scales 1,2 and 3* (2nd ed.). Champaign, II.: Institute for Personality and Ability Test.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Routledge.

Davis, J. P., Bretfelean, L. D., Belanova, E., & Thompson, T. (2020). Super-recognisers: Face recognition performance after variable delay intervals. *Applied Cognitive Psychology*, *34*(6), 1350–1368. https://doi.org/10.1002/acp.3712

Davis, J. M., McKone, E., Dennett, H., O'Connor, K. B., O'Kearney, R., & Palermo, R. (2011). Individual differences in the ability to recognise facial identity are associated with social anxiety. *PloS One*, *6*(12), e28800. https://doi.org/10.1371/journal.pone.0028800

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*(4), 576–585. https://doi.org/10.1016/j.neuropsychologia.2005.07.001

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Field, A. (2011). *Discovering statistics using SPSS: (and sex and drugs and rock 'n' roll)* (3. ed., reprinted.). Los Angeles, Calif.: Sage. Retrieved from http://www.uk.sagepub.com/field3e/main.htm

Fysh, M. C. (2018). Individual differences in the detection, matching and memory of faces. *Cognitive Research: Principles and Implications*, *3*, 20. https://doi.org/10.1186/s41235-018-0111-x

Fysh, M. C., Stacchi, L., & Ramon, M. (2020). Differences between and within individuals, and subprocesses of face cognition: Implications for theory, research and personnel selection. *Royal Society Open Science*, *7*(9), 200233. https://doi.org/10.1098/rsos.200233

Gambarota, F., & Sessa, P. (2019). Visual working memory for faces and facial expressions as a useful "tool" for understanding social and affective cognition. *Frontiers in Psychology*, *10*, 2392. https://doi.org/10.3389/fpsyg.2019.02392

Gignac, G. E., Shankaralingam, M., Walker, K., & Kilpatrick, P. (2016). Short-term memory for faces relates to general intelligence moderately. *Intelligence*, *57*, 96–104. https://doi.org/10.1016/j.intell.2016.05.001

Gregory, R. J. (2014). *Psychological testing: History, principles and applications* (7. ed., global ed.). Boston: Pearson Education.

Haxby, J. V., Hoffman, E. A., & Gobbini, M.I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0

Hayes, A. F., & Coutts, J. J. (2020). Use Omega rather than Cronbach's Alpha for estimating reliability. But… *Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks: SAGE Publications.

Little, T. D. (2013). *The Oxford handbook of quantitative methods in psychology: Vol. 2*: Oxford University Press.

Logie, R., Camos, V., & Cowan, N. (2020). *Working memory*: Oxford University Press.

McCaffery, J. M., Robertson, D. J., Young, A. W., & Burton, A. M. (2018). Individual differences in face identity processing. *Cognitive Research: Principles and Implications*, *3*, 21. https://doi.org/10.1186/s41235-018-0112-9

Megreya, A. M., & Bindemann, M. (2013). Individual differences in personality and face identification. *Journal of Cognitive Psychology*, *25*(1), 30–37. https://doi.org/10.1080/20445911.2012.739153

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865–876. https://doi.org/10.3758/BF03193433

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory. *Intelligence*, *31*(2), 167–193. https://doi.org/10.1016/S0160-2896(02)00115-0

Palermo, R., O'Connor, K. B., Davis, J. M., Irons, J., & McKone, E. (2013). New tests to measure individual differences in matching and labelling facial expressions of emotion, and their association with ability to recognise vocal emotions and facial identity. *PloS One*, *8*(6), e68126. https://doi.org/10.1371/journal.pone.0068126

Palermo, R., & Rhodes, G. (2007). Are you always on my mind? A review of how face perception and attention interact. *Neuropsychologia*, *45*(1), 75–92. https://doi.org/10.1016/j.neuropsychologia.2006.04.025

Palermo, R., Rossion, B., Rhodes, G., Laguesse, R., Tez, T., Hall, B., . . . McKone, E. (2017). Do people have insight into their face recognition abilities? *Quarterly Journal of Experimental Psychology (2006)*, *70*(2), 218–233. https://doi.org/10.1080/17470218.2016.1161058

Petersen, L. A., & Leue, A. (2021). Extraordinary face recognition performance in laboratory and online testing. *Applied Cognitive Psychology*, *3*(2), 22. https://doi.org/10.1002/acp.3805

Peterson, E., & Miller, S. F. (2012). The eyes test as a measure of individual differences: How much of the variance reflects verbal IQ? *Frontiers in Psychology*, *3*, 220. https://doi.org/10.3389/fpsyg.2012.00220

Ramon, M. (2021). Super-Recognizers - a novel diagnostic framework, 70 cases, and guidelines for future work. *Neuropsychologia*, *158*, 107809. https://doi.org/10.1016/j.neuropsychologia.2021.107809

Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology (London, England: 1953)*, *110*(3), 461–479. https://doi.org/10.1111/bjop.12368

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252–257. https://doi.org/10.3758/PBR.16.2.252

Satchell, L. P., Davis, J. P., Julle-Danière, E., Tupper, N., & Marshman, P. (2019). Recognising faces but not traits: Accurate personality judgment from faces is unrelated to superior face memory. *Journal of Research in Personality*, *79*(3), 49–58. https://doi.org/10.1016/j.jrp.2019.02.002

Shah, P., Gaule, A., Sowden, S., Bird, G., & Cook, R. (2015). The 20-item prosopagnosia index (PI20): A self-report instrument for identifying developmental prosopagnosia. *Royal Society Open Science*, *2*(6), 140343. https://doi.org/10.1098/rsos.140343

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(41), 12887–12892. https://doi.org/10.1073/pnas.1421881112

Verhallen, R. J., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Bargary, G., & Mollon, J. D. (2017). General and specific factors in the processing of faces. *Vision Research*, *141*, 217–227. https://doi.org/10.1016/j.visres.2016.12.014

Wechsler, D. (2012). *WAIS-IV. Wechsler Adult Intelligence Scale - Fourth Edition: German Editing by F. Petermann*. Frankfurt, M.: Pearson.

Wilhelm, O., Herzmann, G., Kunina, O., Danthiir, V., Schacht, A., & Sommer, W. (2010). Individual differences in perceiving and recognizing faces-One element of social cognition. *Journal of Personality and Social Psychology*, *99*(3), 530–548. https://doi.org/10.1037/a0019972

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*, 433. https://doi.org/10.3389/fpsyg.2013.00433

Wilmer, J. B. (2017). Individual differences in face recognition: A decade of discovery. *Current Directions in Psychological Science*, *26*(3), 225–230. https://doi.org/10.1177/0963721417710693

Wilmer, J. B., Germine, L. T., & Nakayama, K. (2014). Face recognition: A model specific ability. *Frontiers in Human Neuroscience*, *8*, 769. https://doi.org/10.3389/fnhum.2014.00769

5.3 Anhang zur Studie 3 (Supplement)

The Stirling Face Recognition Scale (SFRS; Bobak et al., 2019) translated into German

Fragebogen zur Wahrnehmung der eigenen Wiedererkennungsfähigkeiten

Dieser Fragebogen enthält 20 Aussagen, die Sie auf einer Skala von 1 (trifft gar nicht zu) bis 5 (trifft voll zu) bewerten sollen. Lesen Sie bitte jede dieser Aussagen aufmerksam durch.

Es gibt keine richtigen oder falschen Antworten. Bitte geben Sie die Antwort an, die am ehesten auf Sie persönlich zutrifft.

Wenn Sie mit der Beantwortung der Aussagen fertig sind, melden Sie sich bei der Testleiterin. Kontrollieren Sie aber bitte vorher, ob Sie auch in jeder Zeile eine Antwort gegeben haben.

1. **Ich weiß genau, wo ich jemanden zum ersten Mal getroffen habe (z.B. bei Arbeitstreffen, Partys etc.)**

☐	☐	☐	☐	☐
Trifft gar nicht zu	*Trifft eher nicht zu*	*neutral*	*trifft eher zu*	*trifft voll zu*

2. **In meiner Schulzeit fiel es mir schwer, meine Klassenkammeraden wiederzuerkennen.**

☐	☐	☐	☐	☐
Trifft gar nicht zu	*Trifft eher nicht zu*	*neutral*	*trifft eher zu*	*trifft voll zu*

3. **Es fällt mir sehr leicht, mir einzelne Gesichter detailreich in Gedanken vorzustellen.**

☐	☐	☐	☐	☐
Trifft gar nicht zu	*Trifft eher nicht zu*	*neutral*	*trifft eher zu*	*trifft voll zu*

4. **Die Angst davor, Gesichter schlecht wiederzuerkennen, hat dazu geführt, dass ich bestimmte soziale oder berufliche Situationen vermeide.**

☐	☐	☐	☐	☐
Trifft gar nicht zu	*Trifft eher nicht zu*	*neutral*	*trifft eher zu*	*trifft voll zu*

5. Ich bin besser als die meisten anderen Personen darin, Namen Gesichtern zuzuordnen.

□ □ □ □ □

Trifft gar nicht zu *Trifft eher nicht zu* *neutral* *trifft eher zu* *trifft voll zu*

6. Ich halte Personen, die ich schon einmal getroffen habe, fälschlicherweise oft für Fremde.

□ □ □ □ □

Trifft gar nicht zu *Trifft eher nicht zu* *neutral* *trifft eher zu* *trifft voll zu*

7. Ich neige dazu, Gesichter sehr schnell wieder zu vergessen.

□ □ □ □ □

Trifft gar nicht zu *Trifft eher nicht zu* *neutral* *trifft eher zu* *trifft voll zu*

8. Mir fällt es sehr leicht, bekannte Personen wiederzuerkennen, wenn ich sie in einem ungewohnten Kontext treffe (z.B. Arbeitskollegen unerwartet im Fitnessstudio treffen).

□ □ □ □ □

Trifft gar nicht zu *Trifft eher nicht zu* *neutral* *trifft eher zu* *trifft voll zu*

9. Gesichter genügen mir, um Personen wiederzuerkennen – ich muss nicht ihre Stimme hören oder ihren gesamten Körper sehen.

□ □ □ □ □

Trifft gar nicht zu *Trifft eher nicht zu* *neutral* *trifft eher zu* *trifft voll zu*

10. Wenn Personen ihren Haarschnitt verändern oder Hüte tragen, habe ich Schwierigkeiten, sie wiederzuerkennen.

□ □ □ □ □

Trifft gar nicht zu *Trifft eher nicht zu* *neutral* *trifft eher zu* *trifft voll zu*

11. Ich bin sehr gut darin, in einer Gruppe von unbekannten Personen miteinander verwandte Personen zu erkennen.

□ □ □ □ □

Trifft gar nicht zu *Trifft eher nicht zu* *neutral* *trifft eher zu* *trifft voll zu*

12. Manchmal muss ich neu kennen gelernte Personen warnen, dass ich schlecht im Wiedererkennen von Gesichtern bin.

☐ ☐ ☐ ☐ ☐

Trifft gar nicht zu *Trifft eher nicht zu* *neutral* *trifft eher zu* *trifft voll zu*

13. Ich erkenne oft Personen wieder, die ich erst kurz zuvor getroffen habe, sie haben jedoch keine Idee, wer ich bin.

☐ ☐ ☐ ☐ ☐

Trifft gar nicht zu *Trifft eher nicht zu* *neutral* *trifft eher zu* *trifft voll zu*

14. Ich muss mich mehr anstrengen als andere Personen, um mir Gesichter zu merken.

☐ ☐ ☐ ☐ ☐

Trifft gar nicht zu *Trifft eher nicht zu* *neutral* *trifft eher zu* *trifft voll zu*

15. Ich vergesse niemals ein Gesicht.

☐ ☐ ☐ ☐ ☐

Trifft gar nicht zu *Trifft eher nicht zu* *neutral* *trifft eher zu* *trifft voll zu*

16. Manchmal fällt es mir schwer, Filmen zu folgen, weil ich Schwierigkeiten habe, Charaktere wiederzuerkennen.

☐ ☐ ☐ ☐ ☐

Trifft gar nicht zu *Trifft eher nicht zu* *neutral* *trifft eher zu* *trifft voll zu*

17. Es fällt mir leicht, meine Familie und Freunde auf Kindheitsfotos wiederzuerkennen.

☐ ☐ ☐ ☐ ☐

Trifft gar nicht zu *Trifft eher nicht zu* *neutral* *trifft eher zu* *trifft voll zu*

18. Es fällt mir schwer, Menschen wiederzuerkennen ohne ihre Stimme zu hören.

☐ ☐ ☐ ☐ ☐

Trifft gar nicht zu *Trifft eher nicht zu* *neutral* *trifft eher zu* *trifft voll zu*

19. Ich erkenne oft jemanden nicht, der mich kennt.

☐	☐	☐	☐	☐
Trifft gar nicht zu	*Trifft eher nicht zu*	*neutral*	*trifft eher zu*	*trifft voll zu*

20. Es fällt mir leicht, eine Person wiederzuerkennen, selbst wenn ich nur Teile ihres Gesichts sehen kann.

☐	☐	☐	☐	☐
Trifft gar nicht zu	*Trifft eher nicht zu*	*neutral*	*trifft eher zu*	*trifft voll zu*

6 Zusammenfassende Diskussion

Ziel dieser Dissertation war es, Einflussfaktoren auf die interindividuelle Gesichtserkennungsleistung im CFMT+ und im GFMT-S zu untersuchen sowie die Befunde zur psychometrischen Qualität der beiden Tests zu erweitern. In drei Studien wurden neue Erkenntnisse über modulierende Faktoren (demografisch, methodisch, kognitiv) sowie zur Reliabilität und Validität des CFMT+ und des GFMT-S gewonnen.

In **Studie 1** wurde gezeigt, dass bei Onlinepräsentation im Vergleich zur Laborpräsentation signifikant höhere Testleistungen im CFMT+ erreicht werden. Ebenso erreichten Proband:innen im Alter zwischen 26 und 35 Jahren im Vergleich zu jüngeren und älteren Proband:innen sowie weibliche Probandinnen im Vergleich zu männlichen Probanden signifikant höhere Testleistungen im CFMT+. Im GFMT-S wurden signifikant höhere Testleistungen nur in der Onlinestichprobe erreicht. Außerdem erreichten Super-Recognizer (überdurchschnittliche Leistungen im CFMT+, normalisierte *z*-Werte) signifikant höhere Testwerte im GFMT-S (normalisierte *z*-Werte). Cronbach's Alpha wurde in der Labor- und Onlinestichprobe berechnet. Mit $N > 300$ war in der Onlinestichprobe eine sehr robuste Berechnung des Reliabilitätskoeffizienten möglich (Charter, 1999). Für den CFMT+ kann Cronbach's Alpha in beiden Stichproben als sehr gut bewertet werden, beim GFMT-S erreichten die Koeffizienten fragwürdige bis akzeptable Werte. Außerdem wiesen die Itemschwierigkeiten im GFMT-S auf zahlreiche leichte Items hin, sodass viele Proband:innen eine sehr hohe Punktzahl im GFMT-S erreichen konnten und ein Deckeneffekt der Testleistung möglich ist. Darüber hinaus wurden erstmals zwei Transformationen auf die Rohwerte im CFMT+ und GFMT-S angewendet (normalisierte *z*-Werte), die die Berechnung von T-Werten und die Veröffentlichung von Normtabellen ermöglicht haben. Mit den Normtabellen ist nun eine standardisierte, einheitliche Einordung der Testleistungen im CFMT+ und im GFMT-S möglich.

Im Fokus von **Studie 2** standen verschiedene Koeffizienten zur Beurteilung der Reliabilität des CFMT+ und des GFMT-S. In einem Test-Retest-Studiendesign zeigte der CFMT+ eine gute Test-Retest-Reliabilität und eine sehr gute interne Konsistenz in Bezug auf die meisten Koeffizienten. Damit ergänzt Studie 2 die Angabe von Cronbach‘s Alpha in Studie 1 durch eine detaillierte Untersuchung der internen Konsistenz mit verschiedenen Koeffizienten. Die Test-Retest-Reliabilität eines diagnostischen Inventars sollte erst interpretiert werden, wenn die interne Konsistenz als gegeben angesehen werden kann (Beauducel & Leue, 2014). Geringe mittlere Inter-Item-Korrelationen (MICs) weisen darauf hin, dass das heterogene Stimulusmaterial im CFMT+ die Berechnung von Cronbach’s Alpha beeinflussen könnte. Darüber hinaus weisen signifikante Mittelwertsunterschiede zwischen den beiden Messzeitpunkten auf Übungseffekte beim CFMT+ hin. Im Gegensatz zum CFMT+ zeigte der GFMT-S keine zufriedenstellende Test-Retest-Reliabilität oder interne Konsistenz.

In **Studie 3** wurde gezeigt, dass die Leistungen in Aufgaben der Aufmerksamkeit, Verarbeitungsgeschwindigkeit und Arbeitsgedächtnis zwar nur gering und nicht signifikant mit den Testleistungen im CFMT+ oder im GFMT-S korrelieren, jedoch in der Summe auch einen Teil zur Varianzaufklärung beitragen. Alle Ergebnisse in Studie 3 sind allerdings aufgrund der angestrebten Datennacherhebung als vorläufig zu betrachten und sollten vorsichtig interpretiert werden (N = 55). Das Testergebnis vom verbalen Arbeitsgedächtnis zeigte insgesamt die höchsten positiven Korrelationen zur Gesichtserkennungstestleistung. Die Verarbeitungsgeschwindigkeit sowie die meisten Maße der Aufmerksamkeit zeigten dagegen negative Korrelationen zur Leistung im CFMT+ und im GFMT-S. Darüber hinaus wurde in Übereinstimmung mit der Literatur ein mittlerer positiver Zusammenhang zwischen der Gesichtserkennungsleistung in beiden Tests (T-Werten) und der selbsteingeschätzten Gesichtserkennungsfähigkeit in einem ins Deutsche übersetzen Fragebogen (SFRS, z-standardisierte Werte) repliziert.

In Abbildung 4 sind die zuvor genannten zentralen Ergebnisse zur Beantwortung der Forschungsfragen der drei Studien zusammenfassend dargestellt (siehe Forschungsfragen in Abschnitten 3.1, 4.1 und 5.1). Im Folgenden werden in Bezug auf diese Ergebnisse die psychometrische Qualität des CFMT+ und des GFMT-S zusammenfassend bewertet und diskutiert. Anschließend werden die drei Studien dieser Arbeit kritisch betrachtet sowie werden abschließend Implikationen und ein Ausblick auf die weitere Forschung beschrieben.

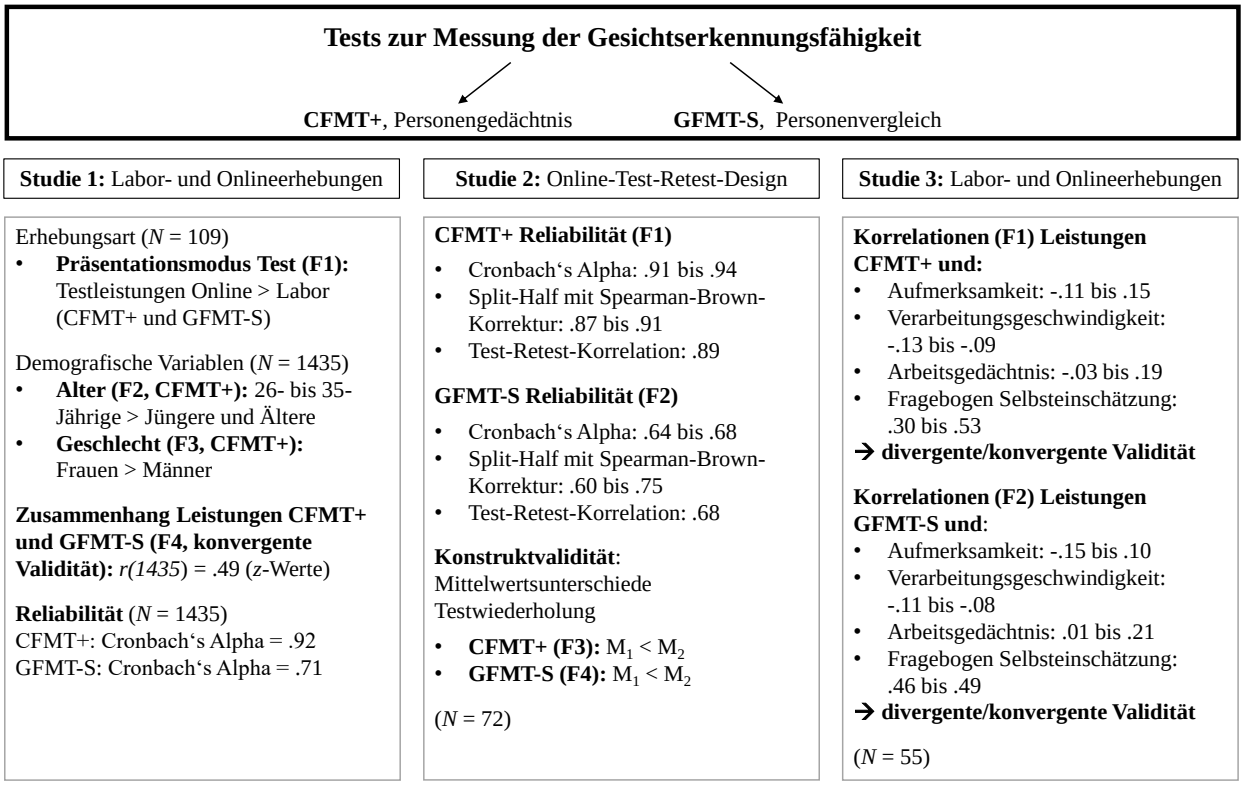


Abbildung 4. Zusammenfassung der zentralen Ergebnisse zur Beantwortung der Forschungsfragen (F) der drei Studien.

6.1 Bewertung der psychometrischen Qualität des CFMT+

Zur Einschätzung der Reliabilität des CFMT+ wurden in den Studien 1 und 2 verschiedene Koeffizienten berechnet: Cronbach’s Alpha (.90 bis .94), Split-Half-Koeffizienten mit Spearman-Brown-Korrektur (.87 bis .91), die Test-Retest-Korrelation (.89) und die mittleren Inter-Item-Korrelationen (MICs; .12 und .18). Mit Ausnahme der MICs

können alle Reliabilitätswerte als gut bis sehr gut bewertet werden (Bühner, 2011; Gregory, 2014). In der Literatur wurde erst kürzlich ein Cronbach's Alpha Koeffizient für den CFMT+ berichtet, der mit .89 (N = 200; Bate et al., 2021) weitgehend zu den Ergebnissen dieser Arbeit passt. Andere psychometrische Kennwerte zur Reliabilität des CFMT+ wurden bislang in der Literatur nicht berichtet. Die niedrigen MICs zum CFMT+ in Studie 2 lassen sich unter anderem durch Effekte heterogenen Stimulusmaterials erklären, da im CFMT+ Gesichter aus verschiedenen Blickwinkeln und in verschiedener Bildqualität gezeigt werden. Das heißt die Konstruktionsart des CFMT+ könnte die Berechnung der Itemhomogenität einschränken, sodass andere Koeffizienten zur Beurteilung der Reliabilität stärker gewichtet werden könnten. Insgesamt sprechen die Ergebnisse aus den Studien für eine gute Reliabilität des CFMT+. Des Weiteren sprechen die Befunde dafür, dass die Anwendung von T-Normwerten für den CFMT+ empfohlen werden kann, da die Anwendung von Feinnormen nur bei reliablen Leistungstests empfohlen wird (Moosbrugger & Kelava, 2020). Die Reliabilitätskoeffizienten im Sinne von Cronbach's Alpha sind jedoch nicht hinreichend, um zu beurteilen, ob das intendierte Konstrukt Gesichtserkennungsfähigkeit auch tatsächlich gemessen wird. Daher ist die Untersuchung der Validität des CFMT+ ebenso entscheidend.

Zur Untersuchung der Konstruktvalidität des CFMT+ wurden Gruppenunterschiede in Studie 1 untersucht, die sowohl in Bezug auf Konstruktunterschiede in den Gruppen als auch in Bezug zur Testfairness sowie in Bezug zu Verteilungsunterschieden interpretiert werden können (vgl. Beauducel & Leue, 2014; Cronbach & Meehl, 1955). Es zeigten sich signifikante Gruppenunterschiede im Alter und im Geschlecht, die mit Ergebnissen des kürzeren CFMT (Alter; Susilo et al., 2013) und anderer publizierter Daten zum CFMT+ (Geschlecht; Bobak, Pampoulov et al., 2016) korrespondieren. In Verbindung mit dem kognitiven Prozess der Gesichtsverarbeitung und der sozialen Komponente (Bruce & Young, 1986; Wilhelm et al., 2010) wäre es möglich, dass unterschiedliche Prozesse oder unterschiedliche Ausprägungen

der kognitiven Verarbeitung die Gesichtsverarbeitung bei Männern und Frauen beeinflussen. Geschlechtsspezifisches Stimulusmaterial könnte daher im CFMT+ untersucht werden. Andererseits können signifikante Geschlechtsunterschiede bedeuten, dass das Konstrukt bei Männern und Frauen zwar identisch ausgeprägt ist, der Test jedoch eine Gruppe bevorzugt und damit bezogen auf auswahlrelevante Kriterien unterschiedlich testfair misst. Ob die Testfairness des CFMT+ beeinträchtigt ist, lässt sich zum Beispiel durch „*Differential-Item-Functioning*"-Analysen (DIF-Analysen) oder durch Prüfung der Messinvarianz nach der klassischen Testtheorie oder nach der *Item-Response*-Theorie (IRT) weiter beurteilen (Beauducel & Leue, 2014; Moosbrugger & Kelava, 2020). In DIF-Analysen werden die Itemschwierigkeiten bei Männern und Frauen innerhalb einer Stichprobe verglichen. Sollten Gruppenunterschiede in Verbindung mit anderen Merkmalsausprägungen stehen, können die Unterschiede ebenso ein gewünschter Effekt der Kriteriumsvalidität sein (Moosbrugger & Kelava, 2020). Weitere Forschung ist demnach notwendig, um zu prüfen, ob Konstruktunterschiede in den Gruppen vorliegen oder eine eingeschränkte Testfairness eine Testüberarbeitung notwendig macht. Darüber hinaus kann weitere Forschung dazu beitragen, auch die Frage nach der Notwendigkeit alters- und geschlechtsspezifischer Normen im CFMT+ zu beantworten.

Die Mittelwertsunterschiede zwischen den zwei Messzeitpunkten in Studie 2 können ebenfalls im Kontext der Konstruktvalidität interpretiert werden (Cronbach & Meehl, 1955). Signifikante Mittelwertsunterschiede im CFMT+ weisen auf Übungseffekte zum zweiten Messzeitpunkt hin. Dies würde bedeuten, dass die Messung des Konstrukts zum zweiten Messzeitpunkt verfälscht wurde und man gegebenenfalls auf einen erneuten Testeinsatz verzichten sollte. Weitere Forschung könnte experimentell Übungseffekte sowie deren Stabilität untersuchen.

Abschließend können im Rahmen der Konstruktvalidität die konvergenten und divergenten Korrelationen aus den Studien 1 und 3 bewertet werden. Die Höhe der konvergenten Korrelationen (zum GFMT-S: .49 bis .61; zum SFRS-Gesamtwert: .47) und der divergenten Korrelationen (zu den kognitiven Subprozessen: -.13 bis .19) zur Leistung im CFMT+ können als zufriedenstellend bewertet werden (vgl. Kline, 2005). Die Höhe der konvergenten Korrelationen zwischen den Leistungen im CFMT+ und im GFMT-S sowie dem SFRS-Gesamtwert passen zu anderen Ergebnissen aus der Literatur (zum Beispiel Bobak et al., 2019; Verhallen et al., 2017). Insgesamt sind die Befunde zur Konstruktvalidität als zufriedenstellend zu beschreiben. Es wird jedoch noch weitere Forschung benötigt, um die Konsequenzen aus den Gruppenunterschieden und den möglichen Übungseffekten abschätzen zu können. Zudem sollten die konvergenten/divergenten Korrelationen mit einer größeren Stichprobe in Studie 3 überprüft werden.

Zusammenfassend stützen die Ergebnisse dieser Arbeit also die psychometrische Qualität des CFMT+. Außerdem zeigen die Ergebnisse, dass es wichtig ist, verschiedene psychometrische Kennwerte zu berechnen, zu interpretieren und kritisch zu diskutieren, um sowohl auf Schwächen als auch auf Stärken in der Messung des Konstrukts im CFMT+ hinweisen zu können.

6.2 Bewertung der psychometrischen Qualität des GFMT-S

Zur Einschätzung der Reliabilität des GFMT-S wurden dieselben psychometrischen Kennwerte wie beim CFMT+ berechnet. Die psychometrischen Kennwerte des GFMT-S erreichten in den Studien 1 und 2 jedoch keine zufriedenstellenden Werte für einen Leistungstest (Bühner, 2011, Gregory, 2014): Cronbach’s Alpha (.64 bis .71), Split-Half-Koeffizienten mit Spearman-Brown-Korrektur (.60 bis .75), Test-Retest-Korrelation (.68) und die MICs (.06 und .07). Die oben genannten Reliabilitätskoeffizienten sind als fragwürdig bis akzeptabel zu bewerten (vgl. Bühner, 2011). Damit passen diese Werte nur zum Teil zu den

wenigen psychometrischen Berichten zum GFMT-S. So erreichte der GFMT-S in Übereinstimmung mit Verhallen et al. (2017; $N = 397$) ein Cronbach's Alpha von .71, aber eine niedrigere Test-Retest-Reliabilität als in Stantic et al. (2021; $r(69) = .77$). Die Ergebnisse dieser Arbeit weisen darauf hin, dass die Messung im GFMT-S weniger reliabel ist. Daher sollten aufgrund der GFMT-S-Testwerte keine diagnostischen Entscheidungen zum Beispiel in der Personalauswahl von Polizeibeamt:innen getroffen werden (vgl. DIN 33430; Beauducel & Leue, 2014; DIN, 2016). Außerdem erwiesen sich alle Items im GFMT-S als sehr einfach, sodass eine geringe Trennschärfe im oberen Leistungsbereich ebenfalls die Interpretation der Testwerte einschränkt. Für die Untersuchung der individuellen Gesichtserkennungsfähigkeiten und die psychologische bzw. personalbezogene Diagnostik von Super-Recognizern scheint der GFMT-S damit ungeeignet, sodass auf andere Verfahren zur Messung der Personenvergleichsfähigkeiten mit besseren Reliabilitätskoeffizienten zurückgegriffen werden sollte. Beispiele anderer kürzlich entwickelter Inventare zum Personenvergleich wie der „*Oxford Face Matching Test*" (OFMT) von Stantic et al. (2021) finden sich in Abschnitt 1.4.2 in dieser Arbeit. Zudem wurde 2021 das Stimulusmaterial des GFMT-S überarbeitet und von White et al. als revidierte Version GFMT 2 veröffentlicht, sodass zukünftig die psychometrische Qualität des GFMT 2 untersucht werden könnte. Die Reliabilität des GFMT-S könnte allerdings ausreichend sein, um den Test im Forschungskontext oder als Screening-Test bei einer Stichprobe mit durchschnittlichen Fähigkeiten im Personenvergleich einzusetzen. So wird in Moosbrugger und Kelava (2020; Kapitel 14) berichtet, dass eine Reliabilität über .70 im Allgemeinen als ausreichend für einen Screening-Test angesehen wird. Darüber hinaus sollten die T-Normwerte aus Studie 1 mit Vorsicht verwendet werden, da bei einer geringen Reliabilität nicht die Anwendung von Feinnormen wie T-Werten empfohlen wird (Moosbrugger & Kelava, 2020).

Die geringen Reliabilitätskoeffizienten des GFMT-S beeinflussen auch die Beurteilung der Validitätsargumente. So ist eine exzellente Reliabilität erstrebenswert, um die Testwerte im Rahmen der Validierung als weitgehend messfehlerfrei annehmen zu dürfen (Moosbrugger & Kelava, 2020). Nur wenn die Testwerte zuverlässig gewonnen wurden, haben beispielsweise Korrelationen zu anderen Messverfahren oder anderen Merkmalen eine angemessene Aussagekraft. So zeigten sich zwar keine signifikanten Gruppenunterschiede beim Alter oder Geschlecht, jedoch könnte es sein, dass Gruppenunterschiede mit dem GFMT-S nur nicht messbar sind. Ebenso müssen die konvergenten Korrelationen (zum CFMT+: 49 bis .61; zum SFRS-Gesamtwert: .49) sowie die divergenten Korrelationen (zu den kognitiven Subprozessen: -.15 bis .21) zur Leistung im GFMT-S aus den Studien 1 und 3 mit Vorsicht interpretiert werden. Die Höhe der konvergenten und divergenten Korrelationen entspricht allerdings den Vorgaben aus der Literatur (vgl. Kline, 2005) und sind damit akzeptabel, sofern die Ergebnisse bei einer größeren Stichprobe in Studie 3 vergleichbar bleiben.

Insgesamt stützen die Ergebnisse der drei Studien nur zum Teil die psychometrische Qualität des GFMT-S. Die überwiegend nicht zufriedenstellenden Reliabilitätskoeffizienten des GFMT-S sprechen derzeit gegen einen Einsatz zur psychologischen bzw. psychometrischen Diagnostik individueller Gesichtserkennungsfähigkeiten. Die ersten Validitätsargumente können zwar als zufriedenstellend bewertet werden, müssen jedoch unter Berücksichtigung der Reliabilität und der erweiterbaren Stichprobengröße von Studie 3 aktuell eingeschränkt interpretiert werden.

6.3 Kritische Bewertung der Arbeit

In diesem Kapitel sollen die Schwächen und Stärken der Studien 1, 2 und 3 diskutiert werden. Es wird dabei sowohl auf die Analysemethoden, die Generalisierbarkeit als auch auf die Messmethoden eingegangen.

Für die Analyse der psychometrischen Qualität des CFMT+ und des GFMT-S wurden in allen drei Studien Koeffizienten der klassischen Testtheorie berechnet. Die Koeffizienten nach klassischer Testtheorie (KTT) haben sich vielfach in der Literatur zur Bewertung der Testgüte bewährt, lassen sich empirisch software-basiert berechnen und werden daher häufig verwendet (Beauducel & Leue, 2014; Moosbrugger & Kelava, 2020). Aufgrund dieser Vorteile bietet es sich an, psychometrische Kennwerte nach KTT zum CFMT+ und zum GFMT-S zu berechnen, um diese für die Forschung nutzbar zu machen. Allerdings unterliegt zum Beispiel Cronbach's Alpha auch strengen theoretischen Voraussetzungen, die in der Regel nicht erfüllt sind (Hayes & Coutts, 2020). Daher wären zur Analyse der Reliabilität des CFMT+ und des GFMT-S auch andere Koeffizienten in Betracht gekommen, zum Beispiel McDonald's Omega (siehe Hayes & Coutts, 2020). Die Berechnung von Omega war aufgrund der Datenlage (Variablen mit Varianz von Null) in dieser Arbeit nicht möglich und auch nicht a-priori als psychometrischer Anspruch formuliert (Hayes & Coutts, 2020). Darüber hinaus hätte die psychometrische Qualität der Tests nach der *Item-Response*-Theorie (IRT) analysiert werden können, wie es beispielsweise schon für den kürzeren CFMT realisiert wurde (Cho et al., 2015). Ein Vorteil von IRT-Analysen ist, dass im Analyse-Modell geschätzte Personenparameter (Schätzung der latenten Variable Gesichtserkennungsfähigkeit) in der Itemanalyse bei der Lösungswahrscheinlichkeit berücksichtigt werden, die sich außerdem auf derselben Skala wie die Itemschwierigkeiten abbilden lassen (siehe Kapitel 12 und 19 in Moosbrugger & Kelava, 2020). Daher wäre es zum Vergleich zu dieser Arbeit für die weitere Forschung interessant, die psychometrische Qualität des CFMT+ mithilfe von IRT-Ergebnissen zu bewerten. Neben dem Vergleich von psychometrischen Analysemethoden wäre es außerdem informativ, Gruppenunterschiede in den Testergebnissen des CFMT+ (Studie 1) mit anderen Methoden zu analysieren. So können Effekte der Analysemethode ausgeschlossen werden, wenn andere Methoden zur selben Interpretation kommen. In Bezug auf den CFMT+ könnten beispielsweise

Effekte des Präsentationsmodus bzw. Motivationseffekte mit *bootstrap-resampling*-Methoden zur robusten Stichprobengewinnung analysiert werden (Field, 2011), um äquivalente alters- und geschlechtsrepräsentative Stichproben vergleichen zu können. Trotz dieser alternativen Analysemethoden muss an dieser Stelle hervorgehoben werden, dass es in dieser Arbeit vor allem darum ging, die psychometrische Qualität des CFMT+ und des GFMT-S erstmals ausführlich zu betrachten. Die Verwendung der Koeffizienten nach KTT stellt daher weniger eine Limitation als einen ausstehenden, wichtigen Erkenntnisgewinn dar. Zudem erweitern die vorliegenden Studien dieser Arbeit die psychometrischen Erkenntnisse zum CFMT+ und zum GFMT-S. Darüber hinaus wurden zum ersten Mal die Testergebnisse vom CFMT+ und GFMT-S für eine deutschsprachige Stichprobe untersucht, was eine wichtige Erweiterung der Ergebnisse aus englischsprachigen Stichproben darstellt.

Neben den statistischen Analysemethoden kann auch die Generalisierbarkeit der Stichproben und Ergebnisse in den Studien bewertet werden. So kann in Studie 3 die zu kleine Stichprobe ($N = 55$) kritisiert werden, weil die Aussagekraft der Ergebnisse dadurch eingeschränkt ist. Daher sollte für Studie 3 eine Datennacherhebung angestrebt werden, um die Analysen mit einer größeren Power (vgl. Faul, Erdfelder, Lang, & Buchner, 2007) und auch mit dem Mindest-N für die angestrebten Analyseverfahren durchzuführen ($N > 100$; Field, 2011). Dennoch können aus den vorläufigen Ergebnissen von Studie 3 Forschungsideen abgeleitet werden wie zum Beispiel die Untersuchung weiterer Einflussfaktoren (siehe Diskussion in Abschnitt 5.1). Als Stärken können dagegen die zufriedenstellend großen Stichproben in Studie 1 ($N = 109$) und Studie 2 ($N = 72$) sowie die sehr große Onlinestichprobe von Studie 1 ($N = 1435$) angesehen werden. Insbesondere die Onlinestichprobe in Studie 1 mit $N = 1435$ kann mit einem ausgeglichenen Geschlechtsverhältnis und einer Altersverteilung von 18 bis 77 Jahren als repräsentativ in Bezug auf die Merkmale bewertet werden. So passt die Stichprobenzusammensetzung in Bezug auf diese Merkmale weitgehend zu den Angaben des

statistischen Bundesamtes (Destatis, 2021). Zudem waren alle Bundesländer Deutschlands in der Stichprobe vertreten. Daher wurde auch eine Generalisierbarkeit bisheriger Ergebnisse aus dem englischsprachigen Raum auf den deutschsprachigen Kontext ermöglicht. Außerdem konnten aufgrund der großen Onlinestichprobe in Studie 1 ($N > 300$; Bühner, 2011; Charter, 1999) zum ersten Mal für den CFMT+ und den GFMT-S T-Normen berechnet werden, die in der weiteren Anwendung der Tests nützlich sein werden. Eine Überprüfung der Normwerte wäre zudem mit im Rahmen dieser Arbeit zusätzlich erhobenen Daten zum CFMT+ und zum GFMT-S möglich. Nach Auswertung der Daten für die Publikation von Studie 1 lief die Onlinedatenerhebung zum CFMT+ und zum GFMT-S weiter, um einen Probandenpool für weitere Studien aufzubauen. So konnten unter anderem die Proband:innen für die Studien 2 und 3 sowie für einige im Rahmen des Projekts gemeinsam mit Prof. Dr. Anja Leue betreute Diplomarbeiten rekrutiert werden (Geyik, 2020; Reiß, 2018; Voß, 2019). Innerhalb von vier Jahren haben 7769 Personen den CFMT+ und den GFMT-S vollständig online in SoSci-Survey bearbeitet. Davon stehen noch ca. 4000 Testdaten für weitere Analysen zur Verfügung. Beispielsweise könnten diese Daten genutzt werden, um die zuvor angesprochenen anderen statistischen Analyseverfahren einzusetzen (unter anderem IRT-Analysen).

Zuletzt werden ebenso die Messmethoden und Herangehensweisen kritisch betrachtet. Die Studien dieser Arbeit wurden überwiegend online durchgeführt. Jedoch können bei Onlinedatenerhebungen weder die Störfaktoren noch die Durchführungsbedingungen kontrolliert werden. Um die Durchführungsbedingungen einschätzen zu können, mussten die Proband:innen in allen drei Studien viele Kontrollfragen beantworten. Auch wenn die Einschätzung der Datenqualität damit von der Aufrichtigkeit der Proband:innen abhing, wurden so nur Daten analysiert, die die vorher festgelegten a-priori Kriterien zur Datenqualität erfüllten. Die Onlinedurchführung wurde trotz der eingeschränkten Kontrollierbarkeit aufgrund der umfangreicheren bzw. repräsentativeren Rekrutierbarkeit von Proband:innen als Messmethode

gewählt, um die Chance auf Identifikation von Super-Recognizern zu erhöhen. Unter der Annahme die Gesichtserkennungsfähigkeit unterliegt einer Normalverteilung, gibt es nur ca. 2% Super-Recognizer in einer Population (Ramon et al., 2019; Russell et al., 2009). Zugleich gibt es seit 2020 durch die COVID-19-Pandemie und die Reduzierung von persönlichen Kontakten (Robert Koch Institut, 2021) einen erhöhten Durchführungsbedarf an Onlinestudien, sodass insbesondere die Untersuchung des Effekts des Präsentationsmodus in Studie 1 eine aktuelle Forschungsfrage darstellt. Insgesamt überwiegen daher die Vorteile der Onlinedatenerhebung in der Erforschung von Super-Recognizern und die Onlinedatenerhebung kann als Messmethode positiv bewertet werden.

6.4 Implikationen und Ausblick

Die Ziele dieser Dissertation konnten durch die Studien 1, 2 und 3 erreicht werden, indem neue Befunde zu Einflussfaktoren gewonnen wurden und die psychometrische Qualität des CFMT+ sowie des GFMT-S geprüft wurde. Die Ergebnisse bringen einige Implikationen für die Gesichtserkennungsforschung und die polizeiliche Praxis mit sich, auf die im Folgenden eingegangen wird.

Die Ergebnisse dieser Arbeit haben einen hohen Forschungsbedarf an Einflussfaktoren auf die Gesichtserkennungsleistung aufgezeigt. Neue Forschungsfragen kamen dabei sowohl in Studie 1 in Bezug auf die signifikanten Gruppenunterschiede (Präsentationsmodus, Alter und Geschlecht) als auch bei den modulierenden kognitiven Faktoren in Studie 3 auf. So sollten beispielsweise die Motivationseffekte von Gesichtserkennungstests in Onlinestudien untersucht werden. Gerade in der Super-Recognizer-Forschung werden Forschungseinheiten von motivierten Proband:innen kontaktiert, um an Tests teilzunehmen, da die Proband:innen selbst berichten, überdurchschnittliche Gesichtserkennungsfähigkeiten zu haben. Es bleibt abzuwarten, ob Super-Recognizer eine signifikant höhere subjektive Einschätzungsfähigkeit der eigenen Gesichtserkennungsleistungen aufzeigen als Kontrollproband:innen (Personen mit

durchschnittlichen Gesichtserkennungsfähigkeiten). So besteht eine Verbindung zwischen der Selbsteinschätzungsfähigkeit und der Testleistung im CFMT+, wie die signifikante mittlere Korrelation zwischen den T-Werten im CFMT+ und dem *z*-standardisierten Gesamtwert des SFRS (Fragebogen zur selbst eingeschätzten Gesichtserkennungsfähigkeit) in Studie 3 trotz der eingeschränkten Stichprobengröße gezeigt hat. Weitere Forschung sollte untersuchen, wie die Selbsteinschätzungsfähigkeit mit der Testleistung interagiert. Sollte in der Gruppe der Super-Recognizer ein *Cut-off*-Wert für das Fragebogenergebnis im SFRS gefunden werden, wäre es denkbar, eine Testbatterie zur psychologischen Diagnostik der individuellen Gesichtserkennungsfähigkeiten durch einen Screening-Fragebogen zu ergänzen.

In Studie 3 wurde neben der Selbsteinschätzungsfähigkeit auch der Einfluss von kognitiven Subprozessen auf die Testleistung im CFMT+ und im GFMT-S untersucht. Hier zeigten insbesondere die negativen Korrelationen zwischen der Gesichtserkennungsleistung und den Leistungen in den WAIS-IV Untertests zur Verarbeitungsgeschwindigkeit, dass die Art der Informationsverarbeitung weiter untersucht werden sollte. Es ist vorstellbar, dass Super-Recognizer länger für Entscheidungen bei Items in Gesichtserkennungstests benötigen und nur methodisch bedingt eine niedrige Verarbeitungsgeschwindigkeit aufzeigen. Üblicherweise wird die Verarbeitungsgeschwindigkeit als Konstrukt unter Zeitdruck gemessen (siehe Konstruktbeschreibung in Wechsler, 2012), während im CFMT+ und im GFMT-S kein Zeitdruck zur Lösung der Items besteht. Es wäre zudem möglich, dass andere kognitive Prozesse im Zusammenhang mit der Gesichtserkennungsleistung stehen. Nach neuen Studien von Belanova et al. (2021) und Dunn et al. (2021) scheint insbesondere der holistische Wahrnehmungsprozess von Gesichtern und Gesichtsmerkmalen ein Anhaltspunkt für die unterschiedlichen Leistungen zu sein. Daher könnte es vielversprechend sein, den Informationsverarbeitungsprozess im Rahmen der individuellen Gesichtserkennungs-fähigkeiten näher zu untersuchen. Mit der Entdeckung von Super-Recognizern wurde ein neuer

Forschungsbereich zu kognitiven Fähigkeiten eröffnet, in dem zurzeit noch viele offene Forschungsfragen existieren. Die weitere Forschung wird zeigen, welche Fähigkeiten Super-Recognizer tatsächlich besitzen, wie die Fähigkeiten in der Praxis genutzt werden können und warum es die individuellen Fähigkeitsunterschiede gibt.

Durch die Ergebnisse dieser Arbeit wurde deutlich, dass es in einem Fachbereich wie der Gesichtserkennungsforschung wichtig ist, die psychometrische Qualität von diagnostischen Inventaren zu überprüfen, da nicht alle Inventare zur Messung individueller Gesichtserkennungsfähigkeiten geeignet sind. So wurde am Beispiel des GFMT-S erkennbar, dass ein häufig eingesetzter Test mit hoher Augenscheinvalidität keine ausreichende psychometrische Qualität aufweisen muss. Ohne Überprüfung der psychometrischen Qualität eines Tests besteht die Gefahr, dass Aussagen aus nicht reliablen Testergebnissen abgeleitet werden, die mit einem reliablen Test nicht haltbar sind. Bislang wurden viele Gesichtserkennungstests nicht ausreichend psychometrisch überprüft. Dies ist jedoch gerade in Anbetracht des steigenden Interesses am Einsatz von Super-Recognizern in verschiedenen Anwendungsbereichen sehr wichtig. Beispielsweise könnte der CFMT+ nach Untersuchung der Kriteriumsvalidität zusammen mit anderen geprüften Tests in der Personalauswahl bei der Polizei eingesetzt werden. Ob Tests wie der CFMT+ mit idealen, im Labor erzeugten Stimuli mit den Gesichtserkennungsfähigkeiten im realen Leben (unter weniger idealen Licht- oder Perspektivbedingungen) zusammenhängen, wird die weitere Forschung zeigen müssen. Der GFMT-S scheint aufgrund der Studienergebnisse bisher nicht für den Einsatz in der Personalauswahl geeignet. Um eine psychologische, psychometrisch-basierte Testbatterie für die Klassifizierung interindividueller Gesichtserkennungsunterschiede mit allen Konstruktfacetten zu gestalten, benötigt es noch mehr psychometrisch geprüfte Gesichtserkennungstests. So würde sich als Alternative zum GFMT-S beispielsweise der 2021 veröffentlichte GFMT 2 (White et al., 2021) zur weiteren Prüfung anbieten, da dieser einige

Kritikpunkte des GFMT-S aufgegriffen hat (unter anderem mehr Variation in den Stimuli). Falls sich unter den vorhandenen Gesichtserkennungstests nicht genügend Tests mit einer zufriedenstellenden psychometrischen Qualität finden lassen, müssten weitere Tests entwickelt werden, um das nomologische Netzwerk der psychologischen, diagnostischen Inventare zur Gesichtserkennungsfähigkeit zu optimieren.

Aufgrund der überwiegend computerbasierten Testungen könnten adaptive Inventare entwickelt werden (Kapitel 20 in Moosbrugger & Kelava, 2020). Adaptive Inventare berücksichtigen die zwischenzeitlich erbrachte Testleistung in der weiteren Auswahl der Items. Beispielsweise könnte nach 10 Items die erreichte Punktzahl entscheiden, ob nun leichtere oder schwerere Items angezeigt werden. So ist es möglich, ein breites Spektrum an Gesichtserkennungsfähigkeiten zu erfassen. Dies wäre zudem eine besonders ökonomische Inventarkonstruktionsart, die in einer kurzen Bearbeitungszeit Fähigkeitsunterschiede sensitiv erfassen kann.

6.5 Abschließende Worte

Diese Dissertation beschäftigte sich mit den individuellen Gesichtserkennungsfähigkeiten. In Übereinstimmung mit der Literatur wurde gezeigt, wie individuell ausgeprägt die Gesichtserkennungsfähigkeiten sind und dass sowohl demografische, methodische als auch kognitive Faktoren die Leistungen in Gesichtserkennungstests modulieren. Darüber hinaus wurde durch die Untersuchung der psychometrischen Qualität des CFMT+ und des GFMT-S deutlich, dass häufig eingesetzte Inventare nicht zur Messung individueller Gesichtserkennungsfähigkeiten geeignet sein müssen. So kann nur der CFMT+ nach Prüfung der psychometrischen Qualität zur Klassifizierung individueller Fähigkeiten empfohlen werden, der GFMT-S jedoch nicht. Insbesondere wenn die Wissenschaftler:innen eine Definition „Super-Recognizer“ abgesprochen haben und eine psychometrisch geprüfte Testbatterie zur Klassifizierung der

individuellen Fähigkeiten abgestimmt ist, können die Fähigkeiten der Super-Recognizer sehr nützlich für die Polizei sein. Insgesamt haben die drei durchgeführten empirischen Studien dieser Arbeit daher zur Beantwortung offener Forschungsfragen beigetragen, aber auch weiteren Forschungsbedarf für zukünftige Studien aufgezeigt.

7 Literaturverzeichnis

American Educational Research Association; American Psychological Association; National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews. Neuroscience, 4*(10), 829–839. https://doi.org/10.1038/nrn1201

Baddeley, A., Hitch, G., & Allen, R. (2020). A Multicomponent Model of Working Memory. In A. Baddeley, G. Hitch, & R. Allen (Eds.), *Working Memory* (pp. 10–43). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0002

Bate, S., & Bennetts, R. J. (2014). The rehabilitation of face recognition impairments: A critical review and future directions. *Frontiers in Human Neuroscience, 8*, 491. https://doi.org/10.3389/fnhum.2014.00491

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., . . . Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications, 3*, 22. https://doi.org/10.1186/s41235-018-0116-5

Bate, S., Portch, E., & Mestry, N. (2021). When two fields collide: Identifying "super-recognisers" for neuropsychological and forensic face recognition research. *Quarterly Journal of Experimental Psychology (2006)*, 17470218211027695. https://doi.org/10.1177/17470218211027695

Beauducel, A., & Leue, A. (2014). *Psychologische Diagnostik. Bachelorstudium Psychologie*. Göttingen: Hogrefe Verlag GmbH.

Belanova, E., Davis, J. P., & Thompson, T. (2018). Cognitive and neural markers of super-recognisers' face processing superiority and enhanced cross-age effect. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *108*, 92–111. https://doi.org/10.1016/j.cortex.2018.07.008

Belanova, E., Davis, J. P., & Thompson, T. (2021). The part-whole effect in super-recognisers and typical-range-ability controls. *Vision Research*, *187*, 75–84. https://doi.org/10.1016/j.visres.2021.06.004

Bennett, R. E., & Davier, M. von (2017). *Advancing human assessment*. Cham: Springer International Publishing.

Bernardi, R. A. (1994). Validating research results when Cronbach'S Alpha is below.70: A methodological procedure. *Educational and Psychological Measurement*, *54*(3), 766–775. https://doi.org/10.1177/0013164494054003023

Blank, H., Anwander, A., & Kriegstein, K. von (2011). Direct structural connections between voice- and face-recognition areas. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, *31*(36), 12906–12915. https://doi.org/10.1523/jneurosci.2091-11.2011

Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *82*, 48–62. https://doi.org/10.1016/j.cortex.2016.05.003

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PloS One*, *11*(2), e0148148. https://doi.org/10.1371/journal.pone.0148148

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

Bobak, A. K., Mileva, V. R., & Hancock, P. J. (2019). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology (2006)*, *72*(4), 872–881. https://doi.org/10.1177/1747021818776145

Bobak, A. K., Pampoulov, P., & Bate, S. (2016). Detecting superior face recognition skills in a large sample of young British adults. *Frontiers in Psychology*, *7*, 1378. https://doi.org/10.3389/fpsyg.2016.01378

Bobak, A. K., Parris, B. A., Gregory, N. J., Bennetts, R. J., & Bate, S. (2017). Eye-movement strategies in developmental prosopagnosia and "super" face recognition. *Quarterly Journal of Experimental Psychology (2006)*, *70*(2), 201–217. https://doi.org/10.1080/17470218.2016.1161059

Brickenkamp, R., Schmidt-Atzert, L., & Liepmann, D. (2010). *d2-R–Aufmerksamkeits- und Konzentrationstest. (d2-R-Attention and concentration test)*. Göttingen: Hogrefe.

Brown, C., & Lloyd-Jones, T. J. (2005). Verbal facilitation of face recognition. *Memory & Cognition*, *33*(8), 1442–1456. https://doi.org/10.3758/BF03193377

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207–218. https://doi.org/10.1037/1076-898X.7.3.207

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology (London, England: 1953)*, *77 (Pt 3)*, 305–327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Bruce, V., & Young, A. (2012). *Face Perception*. London: Psychology Press.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte und erw. Aufl.). *PS Psychologie*. München: Pearson Studium.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*(1), 286–291. https://doi.org/10.3758/BRM.42.1.286

Busey, T. A., & Loftus, G. R. (2007). Cognitive science and the law. *Trends in Cognitive Sciences*, *11*(3), 111–117. https://doi.org/10.1016/j.tics.2006.12.004

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews. Neuroscience*, *6*(8), 641–651. https://doi.org/10.1038/nrn1724

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, *21*(4), 559–566. https://doi.org/10.1076/jcen.21.4.559.889

Cho, S.-J., Wilmer, J., Herzmann, G., McGugin, R. W., Fiset, D., van Gulick, A. E., . . . Gauthier, I. (2015). Item response theory analyses of the Cambridge Face Memory Test (CFMT). *Psychological Assessment*, *27*(2), 552–566. https://doi.org/10.1037/pas0000068

Clark, L. A., & Watson, D. (2016). Constructing validity: Basic issues in objective scale development. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research (4th ed.)* (pp. 187–203). Washington: American Psychological Association. https://doi.org/10.1037/14805-012

Correll, J., Ma, D. S., & Davis, J. P. (2021). Perceptual tuning through contact? Contact interacts with perceptual (not memory-based) face-processing ability to predict cross-race recognition. *Journal of Experimental Social Psychology*, *92*(1), 104058. https://doi.org/10.1016/j.jesp.2020.104058

Cronbach, L. J. (1947). Test reliability; its meaning and determination. *Psychometrika*, *12*(1), 1–16. https://doi.org/10.1007/BF02289289

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Davis, J. P., Bretfelean, L. D., Belanova, E., & Thompson, T. (2020). Super-recognisers: Face recognition performance after variable delay intervals. *Applied Cognitive Psychology*, *34*(6), 1350–1368. https://doi.org/10.1002/acp.3712

Davis, J. P., Forrest, C., Treml, F., & Jansari, A. (2018). Identification from CCTV: Assessing police super-recogniser ability to spot faces in a crowd and susceptibility to change blindness. *Applied Cognitive Psychology*, *32*(3), 337–353. https://doi.org/10.1002/acp.3405

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior face recognition ability in police super-recognisers. *Applied cognitive psychology*, *30*(6), 827–840. https://doi.org/10.1002/acp.3260

Destatis (2021). Bevölkerungsstand. Retrieved from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsstand/_inhalt.html;jsessionid=18DF537C6E52E576028B94F9F3EC769C.live731

DIN (2016). *DIN 33430: Anforderungen an berufsbezogene Eignungsdiagnostik*. Berlin: Beuth.

Dolzycka, D., Herzmann, G., Sommer, W., & Wilhelm, O. (2014). Can training enhance face cognition abilities in middle-aged adults? *PloS One, 9*(3), e90249. https://doi.org/10.1371/journal.pone.0090249

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia, 44*(4), 576–585. https://doi.org/10.1016/j.neuropsychologia.2005.07.001

Dunn, J. D., Lima Varela, V. P. de, Nicholls, V. I., Papinutto, M., White, D., & Miellet, S. (2021). *Visual information sampling in super-recognizers*. https://doi.org/10.31234/osf.io/z2k4a

Dunn, J. D., Summersby, S., Towler, A., Davis, J. P., & White, D. (2020). UNSW Face Test: A screening tool for super-recognizers. *PlosOne, 15*(11). https://doi.org/10.1371/journal.pone.0241747

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Field, A. (2011). *Discovering statistics using SPSS: (and sex and drugs and rock 'n' roll)* (3. ed., reprinted.). Los Angeles, Calif.: Sage.

Fysh, M. C. (2018). Individual differences in the detection, matching and memory of faces. *Cognitive Research: Principles and Implications, 3*, 20. https://doi.org/10.1186/s41235-018-0111-x

Fysh, M. C., & Bindemann, M. (2018). The Kent Face Matching Test. *British Journal of Psychology (London, England: 1953)*, *109*(2), 219–231. https://doi.org/10.1111/bjop.12260

Fysh, M. C., Stacchi, L., & Ramon, M. (2020). Differences between and within individuals, and subprocesses of face cognition: Implications for theory, research and personnel selection. *Royal Society Open Science*, *7*(9), 200233. https://doi.org/10.1098/rsos.200233

Gambarota, F., & Sessa, P. (2019). Visual working memory for faces and facial expressions as a useful "tool" for understanding social and affective cognition. *Frontiers in Psychology*, *10*, 2392. https://doi.org/10.3389/fpsyg.2019.02392

George, D., & Mallery, P. (2020). *IBM SPSS statistics 26 step by step: A simple guide and reference* (Sixteenth edition). New York, London: Routledge.

Geyik, S. (2020). *Einfluss eines Kurz-Onlinetrainings auf die Gesichtserkennungs-abgleichfähigkeit.* Christian-Albrechts-Universität, Kiel. [Diplomarbeit]

Gignac, G. E., Shankaralingam, M., Walker, K., & Kilpatrick, P. (2016). Short-term memory for faces relates to general intelligence moderately. *Intelligence*, *57*, 96–104. https://doi.org/10.1016/j.intell.2016.05.001

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

Gregory, R. J. (2014). *Psychological testing: History, principles and applications* (7. ed., global ed.). Boston: Pearson Education.

Haxby, J. V., Hoffman, E. A., & Gobbini, M.I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0

Hayes, A. F., & Coutts, J. J. (2020). Use Omega rather than Cronbach's Alpha for estimating reliability. But… *Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Henderson, Z., Bruce, V., & Burton, A. M. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology*, *15*(4), 445–464. https://doi.org/10.1002/acp.718

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition*, *21*(9-10), 1306–1336. https://doi.org/10.1080/13506285.2013.823140

Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology. Human Perception and Performance*, *22*(4), 986–1004. https://doi.org/10.1037//0096-1523.22.4.986

Hillstrom, A. P., Sauer, J., & Hope, L. (2011). *Training methods for facial image comparison: A Literature Review*. Retrieved from https://eprints.soton.ac.uk/371613/

Jenkins, R. E., Tsermentseli, S., Monks, C. P., Robertson, D. J., Stevenage, S. V., Symons, A. E., & Davis, J. P. (2021). Are super-face-recognisers also super-voice-recognisers? Evidence from cross-modal identification tasks. *Applied Cognitive Psychology*, *35*(3), 590–605. https://doi.org/10.1002/acp.3813

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks: SAGE Publications.

Kramer, R. S. S., Jones, A. L., & Gous, G. (2021). Individual differences in face and voice matching abilities: The relationship between accuracy and consistency. *Applied Cognitive Psychology*, *35*(1), 192–202. https://doi.org/10.1002/acp.3754

Logie, R., Camos, V., & Cowan, N. (2020). *Working memory*: Oxford University Press.

McCaffery, J. M., Robertson, D. J., Young, A. W., & Burton, A. M. (2018). Individual differences in face identity processing. *Cognitive Research: Principles and Implications*, *3*, 21. https://doi.org/10.1186/s41235-018-0112-9

McKone, E., Wan, L., Pidcock, M., Crookes, K., Reynolds, K., Dawel, A., . . . Fiorentini, C. (2019). A critical period for faces: Other-race face recognition is improved by childhood but not adult social contact. *Scientific Reports*, *9*(1), 12820. https://doi.org/10.1038/s41598-019-49202-0

Megreya, A. M., Bindemann, M., & Havard, C. (2011). Sex differences in unfamiliar face identification: Evidence from matching tasks. *Acta Psychologica*, *137*(1), 83–89. https://doi.org/10.1016/j.actpsy.2011.03.003

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865–876. https://doi.org/10.3758/BF03193433

Megreya, A. M., White, D., & Burton, A. M. (2011). The other-race effect does not rely on memory: Evidence from a matching task. *Quarterly Journal of Experimental Psychology (2006)*, *64*(8), 1473–1483. https://doi.org/10.1080/17470218.2011.575228

Meinhardt-Injac, B., Persike, M., & Meinhardt, G. (2014). Holistic processing and reliance on global viewing strategies in older adults' face perception. *Acta Psychologica*, *151*, 155–163. https://doi.org/10.1016/j.actpsy.2014.06.001

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35. https://doi.org/10.1037/1076-8971.7.1.3

Moosbrugger, H., & Kelava, A. (2020). *Testtheorie und Fragebogenkonstruktion*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Murray, E., & Bate, S. (2020). Diagnosing developmental prosopagnosia: Repeat assessment using the Cambridge Face Memory Test. *Royal Society Open Science*, *7*(9), 200884. https://doi.org/10.1098/rsos.200884

Natu, V., & O'Toole, A. J. (2011). The neural processing of familiar and unfamiliar faces: A review and synopsis. *British Journal of Psychology (London, England: 1953)*, *102*(4), 726–747. https://doi.org/10.1111/j.2044-8295.2011.02053.x

Noyes, E., Davis, J. P., Petrov, N., Gray, K. L. H., & Ritchie, K. L. (2021). The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers. *Royal Society Open Science*, *8*(3), 201169. https://doi.org/10.1098/rsos.201169

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory. *Intelligence*, *31*(2), 167–193. https://doi.org/10.1016/S0160-2896(02)00115-0

Østergaard Knudsen, C., Winther Rasmussen, K., & Gerlach, C. (2021). Gender differences in face recognition: The role of holistic processing. *Visual Cognition*, *29*(6), 379–385. https://doi.org/10.1080/13506285.2021.1930312

Palermo, R., O'Connor, K. B., Davis, J. M., Irons, J., & McKone, E. (2013). New tests to measure individual differences in matching and labelling facial expressions of emotion, and their association with ability to recognise vocal emotions and facial identity. *PloS One*, *8*(6), e68126. https://doi.org/10.1371/journal.pone.0068126

Palermo, R., & Rhodes, G. (2007). Are you always on my mind? A review of how face perception and attention interact. *Neuropsychologia*, *45*(1), 75–92. https://doi.org/10.1016/j.neuropsychologia.2006.04.025

Petersen, L. A., & Leue, A. (2021). Extraordinary face recognition performance in laboratory and online testing. *Applied Cognitive Psychology*, *3*(2), 22. https://doi.org/10.1002/acp.3805

Pointner, N. (2021, August 25). Super-Recognizer bei der Polizei: Sie kenne ich doch. *Spiegel*. Retrieved from https://www.spiegel.de/karriere/super-recogniser-michael-aschenbrenner-erkennt-straftaeter-auf-fotos-a-78057678-811e-4255-a600-24b763572936

Ramon, M. (2021). Super-Recognizers - a novel diagnostic framework, 70 cases, and guidelines for future work. *Neuropsychologia*, *158*, 107809. https://doi.org/10.1016/j.neuropsychologia.2021.107809

Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology (London, England: 1953)*, *110*(3), 461–479. https://doi.org/10.1111/bjop.12368

Reiß, B. (2018). *Super-recognizer: Einfluss verschiedener Persönlichkeitsvariablen auf das Personengedächtnis und Entwicklung eines kulturübergreifenden Personenabgleichtests.* Christian-Albrechts-Universität, Kiel. [Diplomarbeit]

Rezlescu, C., Susilo, T., Wilmer, J. B., & Caramazza, A. (2017). The inversion, part-whole, and composite effects reflect distinct perceptual mechanisms with varied relationships to face recognition. *Journal of Experimental Psychology. Human Perception and Performance*, *43*(12), 1961–1973. https://doi.org/10.1037/xhp0000400

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: A meta-analytic and theoretical review. *Psychological Bulletin*, *138*(1), 146–174. https://doi.org/10.1037/a0025750

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

Robert Koch Institut (2021). COVID-19 (Coronavirus SARS-CoV-2). Retrieved from https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/nCoV.html

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by Metropolitan police super-recognisers. *PloS One*, *11*(2), e0150036. https://doi.org/10.1371/journal.pone.0150036

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252–257. https://doi.org/10.3758/PBR.16.2.252

Satchell, L. P., Davis, J. P., Julle-Danière, E., Tupper, N., & Marshman, P. (2019). Recognising faces but not traits: Accurate personality judgment from faces is unrelated to superior face memory. *Journal of Research in Personality*, *79*(3), 49–58. https://doi.org/10.1016/j.jrp.2019.02.002

Schweinberger, S. R., & Schneider, D. (2014). Wahrnehmung von Personen und soziale Kognition. *Psychologische Rundschau*, *65*(4), 212–226. https://doi.org/10.1026/0033-3042/a000225

Shah, P., Gaule, A., Sowden, S., Bird, G., & Cook, R. (2015). The 20-item prosopagnosia index (PI20): A self-report instrument for identifying developmental prosopagnosia. *Royal Society Open Science*, *2*(6), 140343. https://doi.org/10.1098/rsos.140343

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(41), 12887–12892. https://doi.org/10.1073/pnas.1421881112

Souza, A. C. d., Alexandre, N. M. C., & Guirardello, E. d. B. (2017). Propriedades psicométricas na avaliação de instrumentos: Avaliação da confiabilidade e da validade [Psychometric properties in instruments evaluation of reliability and validity]. *Epidemiologia E Servicos De Saude : Revista Do Sistema Unico De Saude Do Brasil*, *26*(3), 649–659. https://doi.org/10.5123/S1679-49742017000300022

Sporer, S. L., Sauerland, M., & Kocab, K. (2014). Personenidentifizierung. In T. Bliesener, F. Lösel, & G. Köhnken (Eds.), *Lehrbuch der Rechtspsychologie* (1st ed., pp. 156–182). Bern: Verlag Hans Huber.

Stantic, M., Brewer, R., Duchaine, B., Banissy, M. J., Bate, S., Susilo, T., . . . Bird, G. (2021). The Oxford Face Matching Test: A non-biased test of the full range of individual differences in face perception. *Behavior research methods*, *49*(9), 2541. https://doi.org/10.3758/s13428-021-01609-2

Steblay, N. M. (1992). A meta-analytic review of the weapon focus effect. *Law and Human Behavior*, *16*(4), 413–424. https://doi.org/10.1007/BF02352267

Süddeutsche Zeitung (2019, January 24). "Super-Recogniser" helfen bei 200 Fällen. *Süddeutsche Zeitung*. Retrieved from https://www.sueddeutsche.de/muenchen/super-recogniser-polizei-gesichtserkenner-1.4301590

Susilo, T., Germine, L., & Duchaine, B. (2013). Face recognition ability matures late: Evidence from individual differences in young adults. *Journal of Experimental Psychology. Human Perception and Performance*, *39*(5), 1212–1217. https://doi.org/10.1037/a0033469

Thielgen, M. M., Schade, S., & Bosé, C. (2021). Face processing in police service: The relationship between laboratory-based assessment of face processing abilities and performance in a real-world identity matching task. *Cognitive Research: Principles and Implications*, *6*(1), 54. https://doi.org/10.1186/s41235-021-00317-x

Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., & White, D. (2019). Do professional facial image comparison training courses work? *PloS One*, *14*(2), e0211037. https://doi.org/10.1371/journal.pone.0211037

Verhallen, R. J., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Bargary, G., & Mollon, J. D. (2017). General and specific factors in the processing of faces. *Vision Research*, *141*, 217–227. https://doi.org/10.1016/j.visres.2016.12.014

Voß, S. (2019). *Super-recognizer: Differenzierung der Gesichtserkennungsfähigkeiten durch Langzeitgedächtnistests und einen Kurzzeitgedächtnistest mit Perspektivwechsel.* Christian-Albrechts-Universität, Kiel. [Diplomarbeit]

Wechsler, D. (2012). *WAIS-IV. Wechsler Adult Intelligence Scale - Fourth Edition: German Editing by F. Petermann*. Frankfurt, M.: Pearson.

White, D., Guilbert, D., Varela, V. P. L., Jenkins, R., & Burton, A. M. (2021). GFMT2: A psychometric measure of face matching ability. *Behavior Research Methods.* Advance online publication. https://doi.org/10.3758/s13428-021-01638-x

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PloS One*, *9*(8), e103510. https://doi.org/10.1371/journal.pone.0103510

Wilhelm, O., Herzmann, G., Kunina, O., Danthiir, V., Schacht, A., & Sommer, W. (2010). Individual differences in perceiving and recognizing faces-One element of social cognition. *Journal of Personality and Social Psychology*, *99*(3), 530–548. https://doi.org/10.1037/a0019972

Wilhelm, O., Herzmann, G., Kunina, O., & Sommer, W. (2007). Face cognition: A set of distinct mental abilities. *Nature Precedings.* Advance online publication. https://doi.org/10.1038/npre.2007.1385.1

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*, 433. https://doi.org/10.3389/fpsyg.2013.00433

Wilmer, J. B. (2017). Individual differences in face recognition: A decade of discovery. *Current Directions in Psychological Science*, *26*(3), 225–230. https://doi.org/10.1177/0963721417710693

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Gerbasi, M., & Nakayama, K. (2012). Capturing specific abilities as a window into human individuality: The example of face recognition. *Cognitive Neuropsychology*, *29*(5-6), 360–392. https://doi.org/10.1080/02643294.2012.753433

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

Wilmer, J. B., Germine, L. T., & Nakayama, K. (2014). Face recognition: A model specific ability. *Frontiers in Human Neuroscience*, *8*, 769. https://doi.org/10.3389/fnhum.2014.00769

Young, A. W., & Bruce, V. (2011). Understanding person perception. *British Journal of Psychology (London, England: 1953)*, *102*(4), 959–974. https://doi.org/10.1111/j.2044-8295.2011.02045.x