Original Article

Face Memory and Face Matching

Internal Consistency and Test-Retest Reliability for the CFMT+ and the GFMT-S

Lara Aylin Petersen and Anja Leue

Institute of Psychology, Kiel University, Germany

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 α, 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 explaining variations of reliability in face recognition tests.

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

Face perception and face recognition are important components in social and forensic settings (Bruce & Young, 2012). Face memory and face matching abilities focus on face recognition research (Bate et al., 2018; Bobak, Hancock, et al., 2016; Ramon et al., 2019) and face processing research (Verhallen et al., 2017). The Cambridge Face Memory Test Long Form (CFMT+) measuring short-term face memory (Russell et al., 2009) and the Glasgow Face Matching Test Short Form (GFMT-S) measuring face matching (Burton et al., 2010) are frequently used in the 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 α 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 et al., 2014; Hillstrom et al., 2011; White et al., 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 α coefficients should be high even in the assessment context of this study. Cronbach's α integrates the mean inter-item correlations (MICs) and represents the conceptual coherence of all items

$$\alpha = N \times r_{\mathrm{m}}/1 + (N-1) \times r_{\mathrm{m}}, \qquad (1)$$

https://doi.org/10.1027/1614-0001/a000361

where r_m = mean intercorrelation of N items (Bernardi, 1994; Formula 2). Therefore, the higher the MICs in the CFMT+ and the GFMT-S, the more internally consistent the measured construct. Tests that are internally consistent and 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, the 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, just an individual's heritable performance would drive test-retest scores of the CFMT+ and the GFMT-S. Thus, high test-retest reliability coefficients should go along with nonsignificant mean value variations of the CFMT+ and the GFMT-S across measurement points. Therefore, in this study, we wanted to analyze and 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, et al., 2016; Davis et al., 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-month 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) test-retest reliability of $\geq .70$–$.80$ 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, for example, .59 (1 week, $N = 78$; UNSW Face Test; Dunn et al., 2020) and .76 (1 month, $N = 102$; Recognition Memory Test Faces; Bird et al., 2003). Overall, in this study, we investigate the test-retest reliability in conjunction with the internal consistency, 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 fulfill the criteria for sufficient quality of online data were excluded. For example, participants were excluded when they used smartphones 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 had to 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, for example, item difficulties and Cronbach's α, of the adapted online CFMT+ are reported in Petersen and Leue (2021) for laboratory ($N = 109$) and online settings ($N = 1{,}435$; Cronbach's α = .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 the same identity and half with a different identity. The images show female and male faces in frontal position, with a neutral expression, and in gray-scale (for stimuli, see Burton et al., 2010). Participants decided whether the faces

Table 1. Cronbach's α (with CI), MICs, and split-half reliabilities of the CFMT+ and GFMT-S for the first and second measurement points (12 weeks apart)

	CFMT+		GFMT-S	
	First time	Second time	First time	Second time
Cronbach's α	.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

Note. 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 α; 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.

were of the same person or 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, for example, item difficulties and Cronbach's α, of the adapted online GFMT-S conducted in the laboratory (N = 109) and online settings (N = 1,435; e.g., Cronbach's α = .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 = 1,435, 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 section), 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, for example, turning off mobile phones. At the second measurement point, participants were asked to perform the same test conditions as in the first measurement, for example, same time and 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 α is identical with Kuder-Richardson 20 formula when items are scored 0 and 1 (Anselmi et al., 2019; Feldt, 1969), we report the Cronbach's α coefficients. Bühner (2011), as well as George and Mallery (2020), interpreted reliability coefficients (e.g., Cronbach's α) 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 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

The Cambridge Face Memory Test Long Form (CFMT+)

Kolmogorov-Smirnov (K-S) tests showed no normal distribution of the CFMT+ scores for the first and second measurement point (First: K-S(72) = 0.15, $p < .001$; Second: K-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 α and the two split-half reliability coefficients for the first and second measurement points (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).

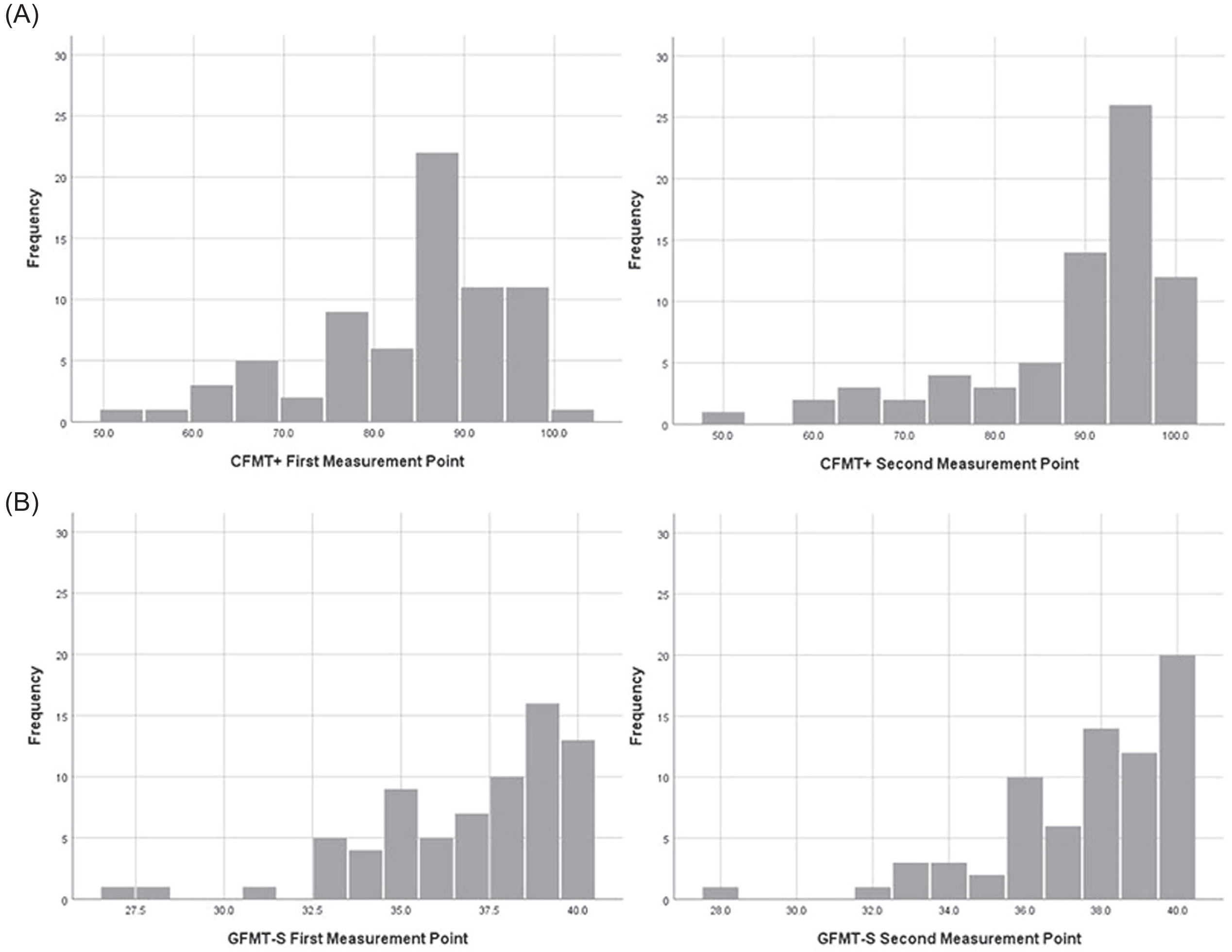

Figure 1. Distribution of the test scores in the CFMT+ (A) and in the GFMT-S (B) for the first measurement point (left) and the second measurement point (right).

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.

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, Mdn = 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, Mdn = 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).

The Glasgow Face Matching Test Short Form (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-S(72) = 0.18, p < .001; Second: K-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 α and the two split-half reliability coefficients (see Table 1) for the first and second measurement points. 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) suggesting 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, Mdn = 38.00, SD = 2.82, range = 27–40, skewness = −1.28) was significantly lower than the mean of the second measurement (M = 37.70, Mdn = 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 α 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 and colleagues (2010; r(42) = .76 for a two-months interval). Further, the CFMT+ has proven reliable because Cronbach's α 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 the 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 the 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 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, for example, because of limited time or disruption. Items of the CFMT+ or the GFMT-S could be preselected based on values of both types of split-half reliability to exclude measurement errors that are due to sequence effects of the items, for example, 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 α 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 α 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 α 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 α coefficient even when the MICs are possibly low due to heterogeneous stimulus material. Moreover, McDonald's ω could be compared to Cronbach's α in further studies (Hayes & Coutts, 2020). One difference between Cronbach's α and ω is that an essential tau equivalence is no prerequisite for ω, but it is for Cronbach's α (Hayes & Coutts, 2020). Therefore, calculating Cronbach's α 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 α and the split-half reliability coefficients of the GFMT-S for the first and second measurement points 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 α 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 zero. The reported test-retest reliability of the GFMT-S corresponds to values of other face matching tests like the KENT face matching test with test-retest reliability of $r(28) = 0.67$ using 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 the construct plus an unsystematic or even systematic error, for example, 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 was 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, 2008).

Conclusions

In conclusion, the CFMT+ showed a satisfactory test-retest reliability and a high to excellent internal consistency (Cronbach's α 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 the 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*. 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*, Article 2714. https://doi.org/10.3389/fpsyg.2019.02714

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*, Article 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*, Article 1378. https://doi.org/10.3389/fpsyg.2016.01378

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

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

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

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). 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), Article 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), Article 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* (16th ed.). Routledge.

Gregory, R. J. (2014). *Psychological testing: History, principles and applications* (7th ed., global ed.). 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. 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*. Information Age Publishing.

Murray, E., & Bate, S. (2020). Diagnosing developmental prosopagnosia. Repeat assessment using the Cambridge Face Memory Test. *Royal Society Open Science, 7*(9), Article 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, 3*(2), 22. 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. D., Alexandre, N. M. C., & de Guirardello, E. B. (2017). Portuguese: Propriedades psicométricas na avaliação de instrumentos: Avaliação da confiabilidade e da validade [Psychometric properties in instruments evaluation of reliability and validity]. *Epidemiologia e Serviços de Saúde: Revista do Sistema Único de Saúde 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., Catmur, C., & 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), Article 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), Article e103510. https://doi.org/10.1371/journal.pone.0103510

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

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

History

Received November 6, 2020
Revision received July 30, 2021
Accepted September 6, 2021
Published online January 27, 2022

Conflict of Interest

We confirm that we have no conflict of interest.

Authorship

Lara A. Petersen und Anja Leue designed the study. L. A. Petersen collected and analyzed the data. L. A. Petersen wrote the manuscript with A. Leue. This study is part of L. A. Petersen's doctoral thesis.

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

Funding

Open access publication enabled by Kiel University.

ORCID

Lara Aylin Petersen
https://orcid.org/0000-0002-1489-0943

Anja Leue
https://orcid.org/0000-0002-2588-5226

Lara Aylin Petersen
Department of Psychology
Kiel University
Olshausenstr. 75
24118 Kiel
Germany
petersen@psychologie.uni-kiel.de