Psychological Test and Assessment Modeling, 2024.66:30-49

DOI: https://doi.org/10.2440/001-0011

Examining the viability of the Continuous Matching Task in mobile assessment compared to laboratory testing

Johann-Christoph Münscher

German Aerospace Center (DLR)

***Abstract**:*
Two measures of attention, the Continuous Matching Task (CMT, measuring alertness), and the Stroop task (measuring selective attention) were applied under two conditions: In the laboratory using a standardized apparatus and in mobile measurements using participant's smart devices. Both are cognitive performance tasks reliant on processing speed. In past research, implementing this type of measurement on mobile devices was called into question and the psychometric quality was assumed to be low. The present study aims to evaluate if the CMT can yield equivalent results from guided laboratory testing and self-administered mobile measurements. The Stroop task results are evaluated in the same way and results of the two tasks are compared. They were implemented identically in both conditions, with only slight modifications to the methods of input. Comparing and analyzing the results revealed that the CMT is not consistent across conditions and prone to age effects on mobile devices. Consequently, it is largely not suited for mobile assessment. The Stroop task showed more consistent measurements, although characteristic shortcomings were also observed. Generally, mobile assessment using response-time-based measurements appear to be problematic when tasks are more technically demanding.

Keywords: Mobile Assessment, Continuous Performance Tasks, Computer Applications, Attention, Reaction Times

Correspondence:

Johann-Christoph Münscher

https://orcid.org/0000-0002-8434-7970

Department of Aviation and Space Psychology, German Aerospace Center (DLR) Institute of Aerospace Medicine.

E-mail: Johann-Christoph.Muenscher@dlr.de

The CMT was developed as a computerized measure of sustained alertness. It employs a unique mode of measurement in which the presentation of stimuli and the recording of responses are continuous – as opposed to presentation at discrete intervals. A target indicator is in continuous linear motion between two points. The motion of this indicator cannot be predicted and must be matched by the participant by moving an analogue slider. As a result, a pronounced level of mental workload is induced while the requirement for higher cognitive processing is low. The task requires alertness and quick responses with little to no cognitive processing. The mental workload can be increased by employing a dual-task paradigm in which two independent instances of the task must be performed at the same time, one with each hand. This effectively constitutes a dual-task paradigm. Another characteristic of the task is that it allows for the real-time adjustment of difficulty. The more the participant succeeds in matching the movement of the indicator, the more difficulty increases and changes in direction occur more frequently; this reduces the predictability and continuity of the motion. In an initial examination, the task was observed to be reliant on reaction speed and thus sustained alertness (Münscher et al. 2023). Furthermore, the adaptive mode of testing in the dual-task paradigm resulted in reliable and conceptually consistent measurements.

During the planning stage of this initial validation, the Covid pandemic emerged and the laboratory design could not be executed as planned. The sample size had to be reduced and a version of the task that could be executed on mobile devices was implemented. Fortunately, the study still commenced with laboratory testing and featured mobile testing. Besides serving as a fallback option in case of lockdowns the inclusion of mobile assessment provided interesting research perspectives: Current smart devices like smartphones can deliver computing power that was limited to desktop and laptop machines only a few years ago. Consequently, they offer an attractive proposition for psychological assessment: Instead of requiring expensive dedicated equipment, researchers and practitioners could make use of their client’s and participant’s devices to perform measurements. In such applications, the smart device is a platform for self-administered testing. Given this scenario and the advantages of testing in this manner, the question arose, whether the mobile implementation of the CMT could provide sensible results when compared to those gathered in the laboratory.

While mobile devices are now regularly used to administer questionnaires, their use in complex designs, such as those involving measurements of response-times is less common. This, among further reasons discussed below, is largely based on the lack of control over the measurement conditions and apparatus in this setting as well as participant behavior. From a methodological standpoint, the self-administered testing using an unknown device in an unstandardized situation constitutes the worst-case scenario as opposed to standardized and controlled laboratory testing. However, there are conflicting accounts on the viability of mobile measurements of cognitive performance and response time.

As Holmlund et al. (2019) pointed out, moving the assessment out of the laboratory and into the hands of individuals comes with specific challenges but also offers worthwhile opportunities. In their application of the Stroop task (Stroop, 1935), they found smartphone-based assessment to be a promising avenue to extend traditional techniques. Illingworth et al. (2015) pointed out that mobile devices put unique and increased demands on the user. They supposed that personality assessments – using questionnaires and other methods that do not rely on response time – are unaffected by the mode of measuring but cognitive assessments are likely influenced negatively.

Regarding the aspects relevant to the smart-phone-based application of the CMT, namely, the measurements of attention with a response-time based performance test that incorporates adaptive testing, a set of promising results were reported in the literature: Koch et al. (2021) showed that measuring sustained attention using the Attention Swiping Task (AST) on mobile devices produced viable results. Similarly, Song et al. (2020) found their smartphone-app-based assessment of cognitive control and executive functioning to be suitable for assessment. Research on the application of adaptive testing in mobile assessment is relatively scarce but an early application of adaptive testing in mobile assessment by Triantafillou et al. (2008) showed promising results.

In many cases, mobile-device-based testing also means uncontrolled testing, which leads to unknown and potentially detrimental conditions under which tests are performed. While laboratory testing is rightfully regarded as superior in this regard, Timmers et al. (2014) observed no differences in task performance between controlled and uncontrolled environments when performing a memory task on the smartphone. Traylor et al. (2020) found no significant influence of the testing environment in a study that included smartphone and laboratory-based assessments of selective attention. Steger et al. (2019) analyzed mobile assessment of intelligence using knowledge tasks, and pointed out that these tasks carry the risk of the participant cheating; a risk they assumed to be much lower in time-dependent tasks. However, with such tasks, the increased demand that is put on the user by interacting with the device may influence the measurement (Illingworth et al., 2015). King et al. (2015) found such influences and observed significant performance differences depending on how cognitive tests were administered. Furthermore, the characteristics of the device, such as its' accuracy in timekeeping can cause distortions. Consequently, the disadvantages of response time tasks on mobile devices are pronounced and Byun et al. (2018) cautioned against such applications. They also reported age effects when performing mobile assessments of reaction times which were also observed by Traylor et al. (2020). However, in a comparison of various tests in mobile and laboratory settings, including cognitive ability tests Martin et al. (2020) found no significant differences in performance. Overall, measurements that rely on response times are likely problematic when they are administered on mobile devices. Should such applications, thus, be avoided outright? Previous finding also indicate that mobile assessment of cognitive performance can be viable and result in useful measurements.

The present study aims to explore and investigate if mobile measurements using the CMT can yield viable results despite the known shortcomings of such measurements. To compare the measurements from the laboratory setting and mobile testing, two tasks were employed: First, the Continuous Matching Task (CMT) (Münscher et al. 2023), a measure of sustained alertness that has not been evaluated in this context. Second, the Stroop task (Stroop, 1935) was applied under both conditions as a measure of selective attention. It showed promising results in previous work by Holmlund et al. (2019) and served as a comparison. The two tasks measure conceptually similar – but not identical – cognitive performances with different methodological approaches. In a computerized implementation, the Stroop tasks can be evaluated using stimulus response-times (RT) while for the CMT the performance over the testing period is evaluated. Details on the tasks and their scoring are elaborated on below. The analyses in the present study were partly based on data collected for the initial validation of the CMT (Münscher et al. 2023). The data on mobile assessments were not included or evaluated in the previous work and are original to this study.

When investigating performance assessments using mobile devices the special circumstances of this mode of testing must be recognized. Specifically, the individual interactions with smart-devices and the associated proficiency in interacting with them. van Deursen et al. (2015) distinguished three types of smartphone use: habitual, process oriented, and social. These types are associated with characteristic patterns of interactions which may influence performance in mobile measurements as the degree to which individuals use these devices varies (Hintze et al., 2017). Furthermore, hardware characteristics are a relevant aspect: Participant's personal devices can vary, particularly in screen size and resolution. When investigating the effects of screen size on performance and workload, Hancock et al. (2015) observed differences only for very small screen sizes and resolutions (320x280 pixels). Smartphone screens are usually larger than this and no substantial effects were expected. Nevertheless, screen dimensions were recognized as a potential confounding variable in the present analyses.

Research Questions and Hypotheses

The primary research question is: Does mobile administration of the CMT, yield results comparable to those gathered in laboratory testing. This inquiry also extends to the Stroop task. It measures a conceptually similar construct (selective attention) and was reported to be promising in mobile applications in the past. Under the optimistic assumption that the two tasks function equally in the testing conditions, the results should be equal. Furthermore, substantial correlations between the laboratory and mobile measurements were expected. As this configuration effectively constitutes a retest, prior results on the test-retest reliability of Stroop tasks were used as guidelines for the expected correlation. Strauss et al. (2005) reported test-retest reliabilities of Stroop color word test RT (n = 28); r_{tt} = .71 for congruent stimuli, r_{tt} = .79 for incongruent stimuli. For interference scores they observed a relationship of r_{tt} = .46. Based on prior reports of age effects for mobile assessment (Byun et al., 2018; Traylor et al., 2020) the same was expected in this study.

For *hypothesis 1* no significant differences between the measurement modes were expected for CMT dual adaptive performance, Stroop RT for congruent and incongruent stimuli, as well as the interference score.

Consequently, *hypothesis 2* was that there is a meaningful correlation in the relevant performance metrics (CMT dual adaptive performance and Stroop RT for congruent and incongruent stimuli, as well as the interference score) between mobile and laboratory scores.

Hypothesis 3 expected a moderation effect of age when predicting laboratory results (CMT dual adaptive performance and Stroop interference score) from mobile testing. Furthermore, no such moderation effects were expected for screen size and types of smartphone use.

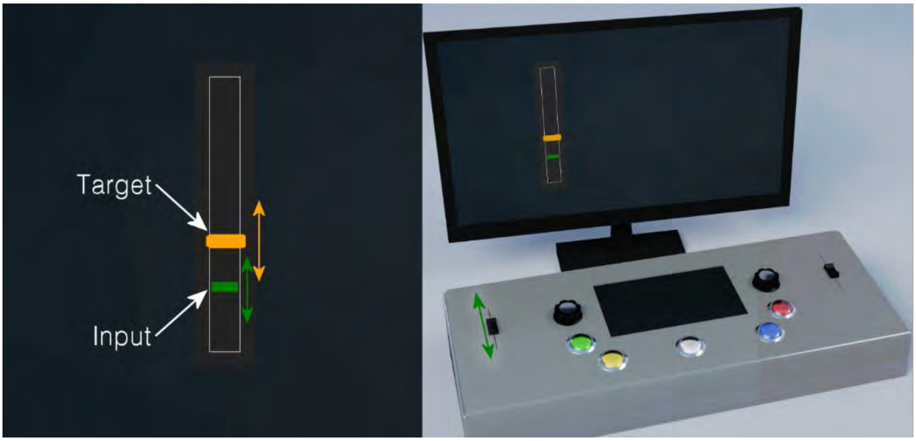
Methods

The sample for the present study consisted of *n* = 125 (59 female (47%), 1 diverse) German participants. Age ranged between 19 and 64 years (*M* = 28, *Mdn* = 24, *SD* = 9.67) and most participants (*n* = 99) were social science students who received course credit for their participation. Additionally, social media postings were used to recruit 26 individuals who were compensated for their participation with a small gift bag containing a USB-drive and university merchandise (pen, textile bag, and a keychain) totaling approximately 5€ in value. Prior to participation, all individuals were informed of the study's intent and gave written informed consent. Participants were instructed to download a specially developed app that included the Stroop and CMT tasks from the distribution platforms Google Play® and Apple® app store. They were then asked to perform the mobile testing either before or after the laboratory session with at least one week of separation. No further instructions were given to the participants in order to allow for in-vivo mobile testing. On average, mobile testing was performed 7.5 days before laboratory assessment (*SD* = 31.3). As the time between testing sessions and the order in which measurements were taken could influence performance, the number of days between sessions was entered as a covariate in the analyses. Participants' mobile devices averaged a screen diagonal of 5.5 inches (*SD* = 0.8). Laboratory testing was performed in the scope of a larger research project and detailed results are published separately (Münscher et al. 2023). All datasets, scripts, and supplementary materials relevant to the following analysis are available in the accompanying OSF repository: https://osf.io/mxkvz/.

Continuous Matching Task (CMT)

The CMT is a continuous performance task that measures sustained alertness (Münscher et al. 2023). The task is fully computerized and makes use of real-time stimuli that are algorithmically generated at the time of testing. A target indicator is continually moving vertically on-screen within a visual boundary. The motion is generated in real-time by a pseudo random number generator that adjusts the target position either up or down at a frequency of 30hz thus generating a continuous motion. Participants control a second (input) indicator within the same boundary by moving an analogue slider. In the laboratory measurements two analogue sliders were present in the the response-box that was used to record responses. The slider must be adjusted to minimize the distance between the target and response indicators; the movement of the target indicator must be matched continuously. See Figure 2 for example stimuli.

Figure 1 CMT stimulus and the laboratory testing apparatus.



Note. Left: An example for CMT stimulus. The target indicator moves within the white boundary and the input indicator must be matched to its' movement. Right: The laboratory apparatus. The response box provides five brightly lit and colored buttons for Stroop testing. The analogue sliders on the left and right sides are used in CMT measurements. The central touchscreen and rotary knobs were not used in this study. The responses were relayed to the testing computer via a low-latency USB connection.

The task can be performed with one hand (main hand or offhand), or with both hands at the same time. In addition to fixed difficulty testing, the CMT offers the option of adaptive testing. In this way of testing, following the target indicator becomes increasingly difficult the better the individual performs. Task difficulty is governed by a difficulty parameter b. It determines both the speed with which the target indicator moves and the frequency of direction changes within the motion. If the indicator reaches either end of the boundary a direction change is forced, in all other instances the direction is changed pseudo-randomly with difficulty b controlling the probability. The difficulty parameter b ranges between 0 and 1 and the task was tuned such that a value of b = 0.5 results in approximately medium difficulty. The performance is evaluated in real-time based on how well the distance between the target and input indicators is minimized. The scoring makes use of a sliding average to compensate for unintentional movements or instances in which the target and response indicators cross each other. Thus, changes in performance need to be relatively persistent to be recognized in the scoring; short-term changes are filtered out.

In the present study the CMT was administered in four configurations: using only the main hand (Single Main hand), using only the offhand (Single Offhand), using both hands with fixed difficulty (b = 0.5, Dual Fixed), and using both hands with adaptive difficulty (Dual Adaptive). All four modes of testing were administered for two minutes each. Depending on the mode of testing different performance indicators must be used. In trials with fixed difficulty, the sliding average distances between target and response are estimators of performance as described above. In adaptive trials, this measure is not informative as the adaption algorithm adjusts the difficulty towards an equilibrium. Instead the adapted difficulty of the task becomes an indicator of participant performance. Both modes of measurement allow for detailed analyses of the development of performance over the time of testing. However, for the present study performance was summarized for each of the two-minute trials. For fixed difficulty testing the average performance across the testing period was calculated. For adaptive testing the integral of the difficulty across the trial was used. Further details on the rationale behind the scoring are elaborated in (Münscher et al. 2023). Furthermore, in trials that employ both hands simultaneously, two independent instances of the CMT are used: one for each hand. To determine the overall performance during these trials, the parallel sum of both hands' performance scores was calculated. Past results indicated that this mode, utilizing both hands with adaptive difficulty, provided the most reliable and valid measurements of sustained alertness (Münscher et al. 2023). Consequently, the analyses that are presented in this study are focused on these measurements; while other performance metrics are reported, the performance indicator from the dual adaptive trial is the main metric of interest.

Stroop Task

The Stroop task is a classic task for measuring selective attention. Participants must read a color word (red, green, blue, yellow) that is displayed on screen, or react to the color in which the word is displayed. Please see Stroop (1935) and specifically Din and Tat Meng (2019) for an overview of this task and computerized implementations.

In the present study, two Stroop trials (color naming, and word reading), were administered with 120 randomized stimuli each. They comprised 48 congruent, 48 incongruent, and 24 neutral stimuli. The stimuli were presented on a screen and responses were recorded using the five brightly lit and colored buttons (red, green, blue yellow, and white for neutral stimuli) featured on the response box.

The Stroop effect describes the interference generated by incongruent stimuli, particularly in the color naming condition. The word is involuntarily read and the content thus interferes with correctly responding to the displayed color. This results in longer response times compared to congruent stimuli. With regards to age effects, Wright (2017) concluded that interference does not vary depending on age, although further research is required. To assess Stroop performance, the response-times (RT; in milliseconds) can be analyzed for all, congruent, incongruent, and for neutral stimuli. This is usually done by averaging the response times to stimuli within a category. Additionally, Golden (1978) recommended a combined score I_G that expresses the individuals' overall ability to suppress interference in all trials. Hit rates and response times were recorded and used to calculate the interference score I_G; higher values indicate a greater ability to process information despite interference.

Laboratory Study and Mobile App implementation

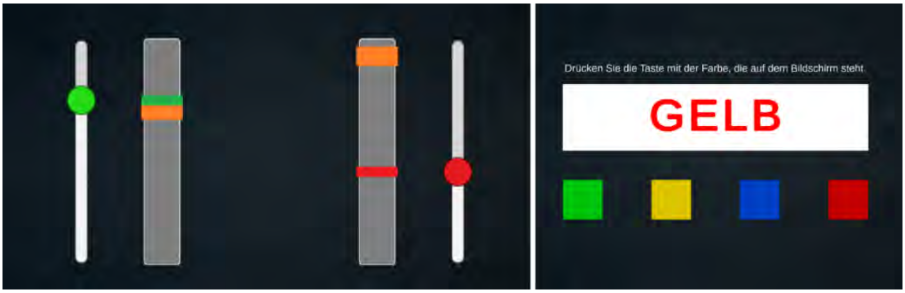
The laboratory measurements were conducted at the Helmut-Schmidt-University / University of the Federal Armed Forces Hamburg over the course of 12 months starting in December 2020. Within the testing session, Stroop and CMT trials were used to gather data for the present analysis. Additionally, the testing session comprised a range of measures that, in the interest of brevity, are not elaborated here. In the laboratory study responses were given using a response box (See Figure 1). For more details on data collection and deployed measures see Münscher et al. (2023).

The two performance tasks, CMT and Stroop, were implemented on mobile devices using the Unity™ game engine. Although it is designed for creating games, it also allows for the implementation of experiments and has been used as platform for experiments in academic research (Brookes et al., 2020; Watson et al., 2018). It showed utility for the present study as it offers a flexible and cost-effective environment for constructing and distributing mobile applications for a range of operating systems. In addition to the individual performance data, the app was also programmed to record the date and the screen dimensions (resolution and pixel density). Further details such as device type, other installed apps, or location were not collected to preserve participant's privacy. Upon completing the testing session, the app transmitted the results of the measurements to an online server; the transmission was encrypted and pseudonymous and contained a unique code identifier that was used to match the laboratory measurements to the mobile ones.

In the current implementation, responses in the app were recorded at a rate of 30hz. For laboratory research, this temporal resolution would be insufficient and dedicated experimental software offers sub-millisecond accuracy. The mobile application does not provide this level of accuracy. However, at the time of writing, no feasible alternative for implementing both stimulus-reaction tasks (Stroop), as well as real-time stimuli (CMT) on both Android™ and IOS™ was available. the Unity™ engine offers a set of technological upsides that cannot be discussed in detail here. In short, its' ease of use, low cost (free), programming language, and intuitive infrastructure to roll out the app to the distribution platforms, makes it an attractive platform for which no comparable alternatives exist. Whilst the Unity™ engine is not the optimal solution for the present study, it was the most promising and pragmatic choice.

The Stroop and CMT tasks were directly adopted to the mobile application, resulting in identical behavior in both scenarios; Stroop stimuli were presented in a different order to avoid repetition. Only the mode of responding was changed to utilize the touchscreens of mobile devices instead of the response-box. Participants were instructed to hold their device horizontally (landscape mode) and the stimulus material was presented in this orientation. When solving the Stroop task, four colored touch fields were displayed under the stimulus material. For the CMT two sliders which participants were instructed to move with their left and right thumb were displayed. The Unity™ engine dynamically scales the onscreen elements based on screen aspect ratio and elements are presented at a consistent relative scale; the scale of the elements thus varies depending on the device. For this implementation of the tasks, the visual elements were arranged so that they would be clearly visible and legible on a variety of common screen sizes with resolutions from 800x480 pixels up to 2960x1404 pixels. See Figure 2 for examples of this implementation in the app.

Figure 2 Screenshots from the implementation of CMT and Stroop tasks in the mobile application.



Drücken Sie die Taste mit der Farbe, die auf dem Bildschirm steht

GELB

Note. Participants were instructed to hold their device horizontally and use their thumbs to respond to the tasks. Left: the CMT in the dual-task configuration. The sliders on the left and right sides (green and red) control the corresponding response indicators. The orange bars (target indicators) move up and down and their motion must be matched. Right: The implementation of the Stroop task in the word reading condition. Currently, an incongruent stimulus (the word "yellow" in red letters) is displayed and the yellow button must be pressed.

Smartphone Usage

As the level of proficiency in handling smart devices, particularly phones, could affect test performance, a measure for assessing the use of smartphones was administered. van Deursen et al. (2015) differentiated three types of smartphone use: habitual, process, and social use. They also developed three short scales (5-point ratings) that result in a score for each of the usage types. *Habitual use* describes an interaction style with the device that is mostly automatic without a specific goal while *process use* is associated with goal-driven behavior (e.g., problem solving, work, looking something up). *Social use* refers to interactions with the device that aim to facilitate social interactions, such as viewing social media posts or messaging.

Statistical Analyses

The measurements in CMT dual adaptive performance, Stroop mean RT for congruent and incongruent stimuli, as well as the interference score were compared using paired-sample t-tests corrected for inequality of variances (Welch-test); the *p* values were holm corrected (Holm, 1979). To test the assumption that scores within each test correlate between modes of testing, the performance metrics of each task were correlated between the measurements (Spearman correlation coefficients with Holm-corrected *p*). The resulting values thus indicate how well the scores from mobile testing correspond to those from laboratory testing. This was performed for both tasks separately, and Fischer-Z tests were performed to test if the absolute observed correlations were significantly smaller than the associations reported by Strauss et al. (2005) ($r = .46$ for Stroop interference scores, $r = .71$ for congruent RT, and $r = .79$ for incongruent RT; $n = 28$). Given these assumed effect sizes, an a-priori power analysis indicated that for $\alpha = .05$ and $\beta = .8$ the required sample size was $n \geq 50$.

Assessing the moderation effect of age on these correlations was performed by multiple linear regression that included age as a moderator. In addition, the analyses included screen size, the number of days between the testing sessions, and the smartphone use scores as moderators. The moderators were included as interaction terms with the mobile performance measurements as predictors and the laboratory measurements as criteria. Again, these analyses were performed for both tasks separately. To control for collinearity the values were mean-centered before regression. In the works of Traylor et al. (2020), and Byun et al. (2018), corrections for age were performed, but no empirical results indicating the strength of the effect were published. Therefore, a small to medium effect of $f^2 = .1$ was assumed and an a-priori power analysis indicated that for $\alpha = .05$ and $\beta = .8$, the required sample size was $n \geq 81$. Calculations were performed in R version 4.3.2 (R Core Team, 2023) using packages “Psych” (Revelle, 2024) and “Interactions” (Long, 2019).

Results

All CMT and Stroop performance metrics showed significant differences between the two testing modes (see Table 1). Performance in the CMT was significantly higher in laboratory testing, as was the Stroop interference score. Furthermore, the responses in Stroop color naming and word reading were significantly quicker in the laboratory setting.

The *n* = 125 Participants reported habitual smartphone use as the most prevalent (*M* = 4.13; *SD* = 0.8) followed by social (*M* = 3.97; *SD* = 0.76) and process use (*M* = 3.18; *SD* = 0.78). Further descriptive statistics for all performance indicators can be found in the supplementary material.

Table 1 Means and standard deviations of CMT and Stroop performance metrics in laboratory and mobile testing.

	Laboratory		Mobile				
	M	*SD*	*M*	*SD*	*t*	*df*	*p*
CMT DA	921.61	98.13	854	121.53	8.29	124	< .01
Stroop I_G	0.4	0.22	0.24	0.14	9.6	123	< .01
Stroop color naming RT	662.5	100.1	860.8	78.9	20.49	124	< .01
Stroop word reading RT	663.8	83.2	794.8	95.4	33.3	124	< .01

Note. Mean comparisons were corrected for inequality of variances (Welch), *p* values were Holm corrected.

Abbreviations: (DA) CMT performance with both hands in adaptive difficulty, (I_G) Golden Interference Score (Golden, 1978), (RT) Mean response times in milliseconds

Regarding hypothesis 1, the assumed parity between the measuring modes was not confirmed. Evidently, laboratory testing yielded higher performance and quicker responses compared to mobile application.

Correlation results

Smartphone use did not exhibit noteworthy associations with both Stroop and CMT performance metrics from mobile measurement. The screen size and the number of days between the testing sessions also did not correlate significantly with the CMT or Stroop metrics. Absolute correlations of CMT scores between laboratory and mobile implementations ranged from |*r*| = .006 to |*r*| = .56. The performances in the dual adaptive trial were significantly correlated with *r* = .53 and this association did not significantly fall behind the expected value of *r* = .71 (*z* = -1.36, *p* = .08); see Table 2 for all correlations. Overall, out of the mobile CMT trials only trials with adaptive difficulty provided sufficiently similar results between the two modes of testing. The absolute correlations of mobile CMT performance indicators with the covariates ranged from |*r*| = .004 to |*r*| = .19, none of which were significant.

Table 2 CMT score correlations between app and laboratory measurement and covariates.

	Laboratory Scores				Covariates				
Mobile Scores	S_o	S_m	DF	DA	In	Δd	Hab	Pr	So
S_o	-.02 -.26; .22	.01 -.19; .21	.03 -.22; .27	-.04 -.29; .22	-.04 -.28; .22	-.17 -.43; .11	-.16 -.42; .12	-.17 -.43; .11	-.04 -.28; .22
S_m	-.03 -.28; .22	-.1 -.36; .18	-.06 -.32; .21	-.01 -.22; .21	.04 -.22; .29	0 -.17; .18	-.07 -.33; .2	-.04 -.29; .22	.04 -.22; .29
DF	.01 -.21; .23	.12 -.16; .38	.06 -.21; .32	-.07 -.33; .2	-.05 -.31; .21	-.09 -.35; .19	-.19 -.45; .09	-.09 -.36; .18	-.05 -.31; .21
DA	-.39* -.61; -.13	-.38* -.59; -.11	-.56* -.73; -.33	**.53*** .29; .7	-.02 -.25; .22	.01 -.22; .24	.05 -.21; .31	-.08 -.34; .2	-.02 -.25; .22

Note. Spearman correlation coefficients and holm corrected CI below. * $p < .05$. (Holm-corrected)

Abbreviations: (S_o) performance of the offhand, (S_m) performance of the main hand, (DF) performance with both hands in fixed difficulty, (DA) performance with both hands in adaptive difficulty, (In) screen diagonal in inches, (Δd) days between testing sessions, (Hab) habitual smartphone use, (Pr) process use, (So) social use. The correlation between DA performances (bold) is relevant to hypothesis 1, it was not significantly weaker than the expected *r* = .71; *p* > .05 (Strauss et al., 2005).

The absolute correlations of Stroop metrics between the measurement modes – mean RT for the stimulus categories and the overall interference score IG – ranged between |r| = .34 and |r| = .69, and were all significant. With regards to hypothesis 2, the correlations of congruent RT in the word reading and color naming conditions did not significantly fall behind the assumed association of r = .71 (word reading: r = .58, z = -1.05, p = .15; color naming: r = .68, z = -0.23, p = .41). The correlations between RT for incongruent stimuli did not match the assumption of r = .79 neither in the word reading condition (r = .54, z = -2.13, p = .002) nor in the color naming condition (r = .58, z = -1.37, p = .08). The correlation between interference scores r = .57 matched its’ expected value of r = .46 (z = -0.7, p = .76); see Table 3 for details.

Table 3 Stroop score correlations between app and laboratory measurement and covariates.

		Laboratory Scores									Covariates				
		Reading				Color naming									
Mobile Scores		t	c	i	n	t	c	i	n	I_G	In	Δd	Hab	Pr	So
Reading	t	.61* .37; .77	.64* .41; .79	.53* .28; .72	.59* .35; .76	.49* .23; .69	.56* .31; .74	.39* .11; .61	.51* .25; .7	-.38* -.6; -.1	-.14 -.39; .13	.2 -.09; .45	-.21 -.47; .08	-.18 -.43; .11	-.07 -.32; .19
	c	.55* .3; .73	**.58*** .33; .75	.49* .23; .69	.54* .29; .73	.46* .19; .66	.51* .26; .7	.36* .08; .59	.49* .23; .69	-.35* -.58; -.07	-.12 -.37; .15	.18 -.1; .44	-.25 -.5; .04	-.17 -.42; .11	-.09 -.34; .17
	i	.61* .37; .77	.63* .4; .79	.54* .29; .72	.59* .35; .76	.47* .2; .67	.54* .28; .72	.37* .09; .59	.48* .21; .68	-.36* -.59; -.08	-.14 -.39; .13	.17 -.11; .43	-.17 -.43; .11	-.17 -.43; .11	-.05 -.29; .18
	n	.52* .26; .7	.58* .33; .75	.42* .15; .64	.49* .23; .69	.47* .21; .67	.54* .28; .72	.38* .11; .61	.49* .23; .68	-.38* -.6; -.1	-.16 -.42; .11	.22 -.07; .48	-.17 -.42; .11	-.16 -.41; .11	-.06 -.3; .19
Color naming	t	.58* .34; .75	.59* .35; .76	.52* .26; .71	.55* .3; .73	.67* .46; .81	.69* .49; .83	.61* .37; .77	.62* .39; .78	-.6* -.77; -.36	-.06 -.3; .19	.22 -.07; .47	-.16 -.41; .12	-.17 -.42; .11	-.03 -.23; .17
	c	.55* .3; .73	.56* .31; .74	.48* .22; .68	.53* .27; .72	.67* .45; .81	**.68*** .47; .82	.6* .36; .77	.62* .39; .78	-.59* -.76; -.35	-.05 -.27; .17	.19 -.09; .45	-.09 -.34; .17	-.14 -.39; .13	0 -.17; .18
	i	.59* .35; .76	.6* .36; .77	.54* .28; .72	.56* .3; .74	.65* .43; .8	.68* .47; .82	.58* .34; .75	.61* .37; .77	-.57* -.75; -.33	-.07 -.31; .19	.23 -.06; .48	-.2 -.46; .09	-.19 -.45; .09	-.06 -.29; .19
	n	.51* .25; .7	.53* .27; .71	.45* .19; .66	.5* .23; .69	.6* .36; .77	.62* .38; .78	.56* .31; .74	.54* .29; .73	-.55* -.73; -.3	-.05 -.28; .18	.19 -.1; .45	-.16 -.41; .12	-.15 -.4; .13	-.05 -.26; .17
I_G		-.59* -.76; -.35	-.6* -.77; -.36	-.54* -.72; -.28	-.56* -.74; -.3	-.65* -.8; - .43	-.68* -.82; -.47	-.58* -.75; -.34	-.61* -.77; -.37	**.57*** .33; .75	.07 -.19; .31	-.23 -.48; .06	.2 -.09; .46	.19 -.09; .45	.06 -.19; .29

Note. Spearman correlation coefficients and Holm-corrected CI. *p < .05. (Holm-corrected) Underlined coefficients mark those that are relevant to testing hypothesis 1, of which those are marked with boldface that do not significantly fall behind the links reported by Strauss et al. (2005); r = .71 for congruent RT, r = .79 for incongruent RT, and r = .46 for interference scores.

Abbreviations: (t) RT for all stimuli, (c) RT for congruent stimuli, (i) RT for incongruent stimuli, (n) RT for neutral stimuli, (I_G) Golden Interference Score (Golden, 1978), (In) screen diagonal in inches, (Δd) days between testing sessions, (Hab) habitual smartphone use, (Pr) process use, (So) social use

Analyzing the correlations of CMT metrics between laboratory and mobile testing indicates that the adaptive CMT is consistent between the measurements compared to the other trials. This is in line with the findings in its' initial validation (Münscher et al. 2023) in which the dual adaptive configuration yielded the best results. The Stroop indicators showed stronger correlations overall, which were significant throughout. Regarding hypothesis 2, the CMT dual adaptive performance showed the expected consistency across the measurements, but only in this performance metric, the remainder of trials did not yield consistent results. Consistency was only partially observed for the Stroop performance. Response times for congruent stimuli correlated significantly and the association reached the expected magnitude whereas the correlations between response times for incongruent stimuli, while also significant, did not. The interference scores I_G were significantly associated in the expected magnitude.

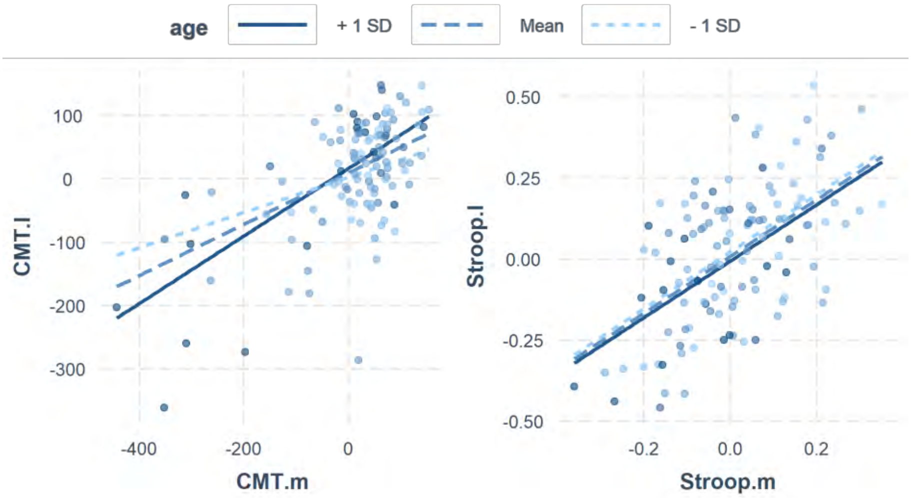
Overall hypothesis 2 can only be partially confirmed for both tasks as their relevant performance indicators showed significant associations between the testing modes. However, the analyses also revealed instances in which inconsistencies were observed. For the CMT only the dual adaptive trial produced consistent results and in the Stroop task the measurements did not reach the expected degree of consistency in all instances.

Moderation analysis

For CMT and Stroop tasks, multiple linear regression analyses predicting laboratory performance from mobile performance were executed. In both, moderations of age, screen size, the number of days between measurements (Δd), and the smartphone use were modeled as interaction terms. The analysis of dual adaptive CMT performance scores resulted in an overall significant model (R = .46, F(13, 103) = 6.76, p < .001). A significant main effect was observed for mobile performance (β = .52, p < .001), but not for age (β = 0.07, p = .465), Δd (β = 0.01, p = .88), screen size (β = 0.1, p = .162), or the types of smartphone use (β = [-.07, .04], p = [.506, .797]). The interaction between mobile dual adaptive performance and age was significant with (β = .31, p = .011), while no interactions were found with Δd (β = -.01, p = -.905), screen size (β = -.03, p = .629), or the types of smartphone use (β = [-.22, .31], p = [.126, .488]).

Slightly different results were observed for Stroop performance. Predicting I_G in the laboratory measurement from mobile measurement with the same moderators resulted in a significant model (R = .37, F(13, 103) = 4.65, p < .001). Mobile performance (β = .56, p < .001) exhibited a significant main effect. Age (β = -.09, p = .442), Δd (β = .06, p = .62), and screen size (β = -0.15, p = .064) did not. Similarly, the types of smartphone use showed no effects (β = [-.04, .01], p = [.698, .895]). The interactions of performance with age (β = -.02, p = .84), Δd (β = .001, p = .99), screen size (β = .08, p = .321), and smartphone use (β = [-.06, .07], p = [.559, .862]) were also not significant. Detailed results for both analyses can be found in the supplementary material. Figure 3 illustrates the moderation of age in both tasks with three age bins (*m*, +1*sd*, -1*sd*).

Figure 3 Plots for CMT and Stroop scores in laboratory and mobile measurement, moderated by age.



Note. Scores are mean-centered. Abbreviations: (CMT.l) CMT laboratory score in the dual adaptive trial, (CMT.m) CMT mobile score in the dual adaptive trial, (Stroop.l) Stroop laboratory interference score I_G, (Stroop.m) Stroop mobile interference score I_G). For both tasks, an association between laboratory and mobile measurements was observed. Age moderated the association of CMT performance between mobile and laboratory testing.

Regarding hypothesis 3, significant age affects were only observed for the CMT while the Stroop interference was not moderated by age. For both tasks, no moderating effects by screen size, the number of days between measurements, and the smartphone use were observed.

Discussion

The present study aimed to investigate how the CMT and the Stroop task perform in laboratory application compared to the administration on mobile devices. Comparisons of the results indicate at least two things. First: The two tasks do not seem to produce results that are independent of the mode of measurement. Second: Both tasks present characteristic drawbacks when employed in mobile assessment.

Comparing the levels of performance between the measurement modes revealed that laboratory testing yielded consistently better performance. While reports by Timmers et al. (2014), Traylor et al. (2020), and Martin et al. (2020) indicated that no substantial differences between the measurement modes were observed, these findings were not replicated in the present study. Responses were given significantly quicker in the Stroop task and the interference score was higher, as was the CMT performance in the dual adaptive trial. These results mirror the findings by King et al. (2015) who also observed significantly better performance in laboratory testing. The exact source of these discrepancies cannot be determined within the scope of this study with resect to the research question it has become clear that neither task delivers equal results in the two measurement modes. Likely, the increased demand that is put in the participant when using mobile devices contributed to the reduced performance (Illingworth et al.,2015). As neither correlations nor moderation effects of screen size and smartphone use types were observed, the performance decrease is likely caused by the input method and the uncontrolled testing environment. Beyond the levels of performance, the consistency of measurements between the measurement modes was of interest in hypothesis 2. Stroop performance was overall strongly linked between laboratory and mobile devices. Using test-retest correlations for Stroop tasks (Strauss et al., 2005) as guidelines, the mobile application of the Stroop task fared well in this study. This is in line with findings by Holmlund et al. (2019) who found the Stroop task to yield usable results in mobile measurements. Some performance indicators exhibited associations that mirrored their previous findings. Namely those of the color naming condition, in which Stroop interference occurs most prominently. Furthermore, the findings by Wright (2017), that interference is independent of age, were replicated here. For the combined interference score I_G a link matching the assumed magnitude was also observed without being moderated by age. While mobile device implementations of tests of attention and cognitive control were reported to be viable (Koch et al., 2021, Song et al., 2020), the same cannot be said for the CMT. The primary CMT score did not exhibit associations as strong as those observed in the Stroop task. Nevertheless, the results were significantly correlated matching the expected magnitude. While the dual adaptive trial yielded sufficiently consistent results, the performance metrics from the remaining trials did not. On the one hand this is in line with previous results in which the dual adaptive trial resulted in the most reliable and valid measurements. On the other hand, this indicates that CMT measurements are largely not consistent between measurement modes, except for the dual adaptive trial. If the CMT is to be deployed on mobile devices, the dual adaptive trials are the only option to yield relatively consistent results but the reduced performance and age effects must be considered.

Concerning the effects of age, the present analysis revealed that CMT performance between mobile and laboratory testing was moderated by age. However, age did not influence the association between laboratory and mobile performance in Stroop trials. This age effect mirrors the findings by Byun et al. (2018) and Traylor et al. (2020) and indicates that applying the CMT in mobile assessment is problematic. Older participants in the lower range of laboratory performance performed worse in mobile testing compared to younger individuals who showed similar performance in laboratory testing. The older participants likely struggled with the mobile implementation and its' use of the touchscreen. Substituting laboratory CMT testing with mobile assessment is therefore not recommended, especially when older participants are recruited. In both tasks, no significant moderation effects of smartphone use, screen size, or the number of days between testing were observed when predicting laboratory from mobile measurements. The independence of screen size supports the conclusion by Hancock et al. (2015) who found only very small screens to be problematic. Types of smartphone use showed no moderating effects and no significant associations with mobile task performance in Stroop or CMT. These findings align with those by Koch et al. (2021), who observed that participants attitude towards technology did not influence measurements.

In summary, these findings indicate that the Stroop task is overall better suited for an application in mobile assessment. However, both tasks have exhibited characteristic shortcomings. These must be considered when deploying mobile assessments.

Contribution

The present study highlights the psychometric qualities of two time-dependent tasks in laboratory and mobile applications. Results align with the conclusion by Byun et al. (2018) and indicate that such assessments using mobile devices can be problematic. Furthermore, the age effects brought up by Byun et al.; Traylor et al. (2018; 2020) were observed in the CMT but not the Stroop task. For technically complex tasks like the CMT mobile assessment cannot be used to substitute laboratory testing without thorough testing and consideration of the target demographic. Therefore, laboratory studies are required for tasks that are technically more advanced than simple stimulus-reaction tasks.

Limitations and Outlook

The primary shortcoming of the present study is its' sample size. A more substantial sample would have allowed for a detailed analysis of method effects using structural equation modeling and assessments of measurement invariance. The current sample of n = 125 was too small to adequately perform such analyses. However, the findings are sufficiently robust to identify the problems that arise in such transfers. Future inquiries should make use of a larger sample size. Furthermore, the trial sessions were only two minutes long. The CMT is a task of sustained alertness, which may return different results for a longer time-on-task. Also, only CMT and Stroop tasks were performed and analyzed. Including other tasks, such as other continuous performance tasks could have enabled more detailed analyses of the origins of the inaccuracies that are introduced by mobile administration. Furthermore, additional retests of mobile measurements could have helped to further highlight the differences between measurement modes. As it stands, this study does not feature a true test-retest. In addition, the present analysis does not include any external or additional criteria to compare against. Here, administering other measures in both laboratory and mobile settings could enable a multi-trait multi- method approach to highlight the effects of measurement mode more precisely.

Conclusion

The present study highlights that the promise of self-administered assessment using reaction-time-based tasks can only be realized with certain tasks. Based on the present findings and prior research (Byun et al., 2018; Illingworth et al., 2015), response-time-based tasks are not universally applicable to mobile testing. Stroop tasks show promise in this regard, while a continuous and more complex measure like the CMT was shown to be limited in this application. An exception may be an application specifically geared towards young participants. The age effects may be less pronounced in these populations so that it or similar tasks could be informative. Although the CMT can generally not be recommended for use in mobile testing, the reliability gained from adaptive testing also became evident. Therefore, adaptive testing could be leveraged to counter similar shortcomings of other tasks. In general, mobile assessment based on time-dependent tasks should, thus, be applied cautiously and with rigorous testing of the tasks and implementations. Measurement equivalence between laboratory and mobile measurements cannot be assumed just because the tasks function technically identically in both scenarios.

References

Brookes, J., Warburton, M., Alghadier, M., Mon-Williams, M., & Mushtaq, F. (2020). Studying human behavior with virtual reality: The Unity Experiment Framework. *Behavior Research Methods*, *52*(2), 455–463. https://doi.org/10.3758/s13428-019-01242-0

Byun, Y. S., Park, S. K., Sakong, J., & Jeon, M. J. (2018). Performance assessment on the Korean Computerized Neurobehavioral Test using a mobile device and a conventional computer: An experimental study. *Annals of Occupational and Environmental Medicine*, *30*, 55. https://doi.org/10.1186/s40557-018-0264-6

Din, N. C., & Tat Meng, E. C. (2019). Computerized Stroop Tests: A Review. *Journal of Psychology & Psychotherapy.* Advance online publication. https://doi.org/10.4172/2161-0487.1000353

Golden, C. J. (1978). *Stroop Color and Word Test: A Manual for Clinical and Experimental Uses*. Stoelting Co.

Hancock, P. A., Sawyer, B. D., & Stafford, S. (2015). The effects of display size on performance. *Ergonomics*, *58*(3), 337–354. https://doi.org/10.1080/00140139.2014.973914

Hintze, D., Hintze, P., Findling, R. D., & Mayrhofer, R. (2017). A Large-Scale, Long-Term Analysis of Mobile Device Usage Characteristics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(2), 1–21. https://doi.org/10.1145/3090078

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70. https://www.jstor.org/stable/4615733.

Holmlund, T. B., Foltz, P. W., Cohen, A. S., Johansen, H. D., Sigurdsen, R., Fugelli, P., Bergsager, D., Cheng, J., Bernstein, J., Rosenfeld, E., & Elvevåg, B. (2019). Moving psychological assessment out of the controlled laboratory setting: Practical challenges. *Psychological Assessment*, *31*(3), 292–303. https://doi.org/10.1037/pas0000647

Illingworth, A. J., Morelli, N. A., Scott, J. C., & Boyd, S. L. (2015). Internet-Based, Unproctored Assessments on Mobile and Non-Mobile Devices: Usage, Measurement Equivalence, and Outcomes. *Journal of Business and Psychology*, *30*(2), 325–343. https://doi.org/10.1007/s10869-014-9363-8

King, D. D., Ryan, A. M., Kantrowitz, T., Grelle, D., & Dainis, A. (2015). Mobile internet testing: An analysis of equivalence, individual differences, and reactions. *International Journal of Selection and Assessment*, 23(4), 382-394.

Koch, M., Möller, C., & Spinath, F. M. (2021). Are You Swiping, or Just Marking? Exploring the Feasibility of Psychological Testing on Mobile Devices. *Psychological Test and Assessment Modeling*, 63(4), 507-524.

Long, J.A. (2019). interactions: Comprehensive, User-Friendly Toolkit for Probing Interactions. R package version 1.1.0, https://cran.r-project.org/package=interactions

Martin, N.R., Capman, J., Boyce, A.S., Morgan, K.E., Gonzalez, M.F., & Adler, S. (2020). New frontiers in cognitive ability testing: working memory. *Journal of Managerial Psychology*, 35, 193-208.

Münscher, J. C., Bürger, M., & Herzberg, P. Y. (2023). The Continuous Matching Task (CMT)–real-time procedural stimulus generation for adaptive testing of attention. *Applied Neuropsychology*: Adult, 30(5), 577-590. https://doi.org/10.1080/23279095.2021.1969399

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Revelle, W. (2024). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. R package version 2.4.1, https://CRAN.R-project.org/package=psych

Song, H., Yi, D.-J., & Park, H.-J. (2020). Validation of a mobile game-based assessment of cognitive control among children and adolescents. *PloS One*, *15*(3), e0230498. https://doi.org/10.1371/journal.pone.0230498

Steger, D., Schroeders, U., & Wilhelm, O. (2019). On the dimensionality of crystallized intelligence: A smartphone-based assessment. *Intelligence*, *72*, 76–85. https://doi.org/10.1016/j.intell.2018.12.002

Steiger, J. H. (1980). Testing Pattern Hypotheses On Correlation Matrices: Alternative Statistics And Some Empirical Results. *Multivariate Behavioral Research*, *15*(3), 335–352. https://doi.org/10.1207/s15327906mbr1503_7

Strauss, G. P., Allen, D. N., Jorgensen, M. L., & Cramer, S. L. (2005). Test-retest reliability of standard and emotional stroop tasks: An investigation of color-word and picture-word versions. *Assessment*, *12*(3), 330–337. https://doi.org/10.1177/1073191105276375

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

Timmers, C., Maeghs, A., Vestjens, M., Bonnemayer, C., Hamers, H., & Blokland, A. (2014). Ambulant cognitive assessment using a smartphone. *Applied Neuropsychology. Adult*, *21*(2), 136–142. https://doi.org/10.1080/09084282.2013.778261

Traylor, Z., Hagen, E., Williams, A., & Arthur, W. (2020). The testing environment as an explanation for unproctored internet-based testing device-type effects. *International Journal of Selection and Assessment.* Advance online publication. https://doi.org/10.1111/ijsa.12315

Triantafillou, E., Georgiadou, E., & Economides, A. A. (2008). The design and evaluation of a computerized adaptive test on mobile devices. *Computers & Education*, *50*(4), 1319–1330. https://doi.org/10.1016/j.compedu.2006.12.005

van Deursen, A. J., Bolle, C. L., Hegner, S. M., & Kommers, P. A. (2015). Modeling habitual and addictive smartphone behavior. *Computers in Human Behavior*, *45*, 411–420. https://doi.org/10.1016/j.chb.2014.12.039

Watson, M. R., Benjamin, V., Christopher, T., Asif, H., & Thilo, W. (2018). *USE: An integrative suite for temporally-precise psychophysical experiments in virtual environments for human, nonhuman, and artificially intelligent agents.* https://doi.org/10.1101/434944

Wright, B. C. (2017). What Stroop tasks can tell us about selective attention from childhood to adulthood. *British Journal of Psychology (London, England : 1953)*, *108*(3), 583–607. https://doi.org/10.1111/bjop.12230