Within-subject reliability, occasion specificity, and validity of fluctuations of the Stroop and Go/No-Go tasks in ecological momentary assessment

Justin Hachenberger[1], Axel Mayer[1], Denny Kerkhoff[1], Friederike Eyssel[1,2], Stefan Fries[1], Tina B. Lonsdorf[1,3], Hilmar Zech[4,5], Lorenz Deserno[4,5], and Sakari Lemola[1]

[1]Department of Psychology, Bielefeld University, Bielefeld, Germany
[2]Center for Cognitive Interaction Technology (CITEC), Bielefeld University, Bielefeld, Germany
[3]Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[4]Department of Psychiatry and Psychotherapy, Technische Universität Dresden, Dresden, Germany
[5]Department of Child and Adolescent Psychiatry, Psychotherapy and Psychosomatics, Center for Mental Health, University Hospital Würzburg, Würzburg, Germany

*Correspondence concerning this article should be addressed to Justin Hachenberger, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany. Email: justin.hachenberger@uni-bielefeld.de

Abstract

Following the (revised) latent state-trait theory the present study investigates within-subject reliability, occasion specificity, common consistency, and construct validity of cognitive control measures in an intensive longitudinal design. These indices were calculated applying dynamic structural equation modelling while accounting for autoregressive effects and trait change. In two studies, participants completed two cognitive control tasks (Stroop and Go/No-Go) and answered questions about goal-pursuit, self-control, executive functions, and situational aspects, multiple times per day. The sample (aged 18–30 years in both studies) consisted of 21 participants (14 females) in the Pilot Study and 70 participants (48 female) in the Main Study. Findings indicate poor within-subject reliability for the Stroop task error rate and reaction time difference between congruent and incongruent trials and moderate to good within-subject reliability for the Go/No-Go task error rate and reaction time. Occasion specificity – the systematic variance accounted for by state residuals – was on a modest level (between 1.4% and 11.1%) for the Stroop error rate and reaction time difference, and on a moderate level (between 16.1% and 37.2% for the Go/No-Go error rate and reaction time) in the two studies. Common consistency – the variance accounted for by latent trait variables – was on a moderate to high level for all of the investigated scores. Indicative of construct validity, the Stroop and Go/No-Go task error rates correlated positively with each other on the within- and between-subject level. Within-subject correlations between task scores and subjective self-control measures were very small and mostly not significant.

Keywords: cognitive control; stroop; go/no-go; ecological momentary assessment; dynamic structural equation modeling; intensive longitudinal design

Introduction

The ability to apply cognitive control over one's actions and to inhibit dominant responses has typically been measured with cognitive control tasks such as the Stroop task (MacLeod, 1991) or the Go/No-Go task (Falkenstein et al., 1999). Recently, researchers have started to embed cognitive control tasks into the context of ecological momentary assessment (EMA) to either measure whether fluctuations in cognitive control are predicted by fluctuations in intrapsychic or contextual variables or whether cognitive control predicts behavioral outcomes (e.g., Zech et al. 2023). First studies indicate that fluctuations in cognitive control measured by EMA can predict changes in real-world behavior. For example, Powell et al. (2017) showed that intra-individual fluctuations of the performance on the Go/No-Go task predicted subsequent snacking behavior in within-subject analyses, whereas no associations of the person mean performance with snacking behavior was observed in between-subject analyses. Moreover, better state-level performance on the Go/No-Go task was associated with reduced negative emotionality (Rónai et al., 2023) and better mood specifically in individuals with higher trait-levels of resilience (Nahum et al., 2023), indicating the potential importance of within-individual fluctuations of cognitive control for positive functioning and real-world behavior. Similarly, within-individual fluctuations in the reaction time measured with the Stop-Signal Task predicted fluctuations in alcohol consumption in heavy drinkers (Jones et al., 2017). Thus, within-individual fluctuations in cognitive control tasks appear to predict within-individual fluctuations in real-world behavior.

Despite growing interest in within-individual fluctuations of cognitive control in the ecological context, existing research on the reliability and validity of cognitive control tasks has mainly focused on the laboratory context. Using only a limited number of measurement occasions this laboratory-based research has shown moderate to excellent levels of (overall)

reliability and modest to moderate correlations between different cognitive control tasks, depending on the specific tasks studied (see e.g., Faßbender et al., 2023). While research in the EMA context is still scarce, existing studies on reliability and validity of a cognitive control task (i.e., the Stop Signal Reaction Time) in the intensive longitudinal design indicated moderate levels of retest-reliability and moderately high factor loadings on a common executive control factor (Zech et al. 2022; Zech et al. 2023). Surprisingly, however, existing research on cognitive control measures in EMA has to our knowledge not reported within-subject variability measures such as within-subject reliability and occasion-specificity. The former – within-subject reliability – quantifies the reliability of the measurement of within-person fluctuations. It is also referred to as the "reliability of change" (Neubauer & Schmiedek, 2020) or the measurement of shared “ups and downs” of indicators across time (Brose et al., 2020). In contrast, the latter – occasion-specificity – measures the proportion of overall variance in the response that is due to systematic situational effects, including situation-person interactions.

Specifically, occasion-specificity is defined by the revised Latent-State-Trait Theory (LST-R; Steyer et al., 2015), an extension of classical test theory. According to LST-R, total variance can be decomposed into “true variance” and “error variance”, and true variance can be further decomposed into “trait-level variance” and “state residual variance”. The proportions of trait-level variance and state residual variance in the total variance are defined as common consistency and occasion specificity, respectively. While common consistency is assumed to be mainly due to genetic and long-standing environmental influences as well as gene-environment interactions (Plomin et al., 1977), occasion specificity is thought to be impacted by situational influences and person by situation interactions (Steyer et al., 1999). Thus, both within-person reliability and occasion specificity quantify aspects of the measurement of within-person fluctuations. However, they measure distinct components:

While within-person reliability measures the reliability of a set of items to capture systematic fluctuations within persons and across time (i.e., measurement occasions) specifically in relation to the within-person variance, occasion-specificity measures the proportion of systematic within-person variance (i.e., situational effects and person-situation interactions) in relation to the total response variance. Reporting of these indices is important to estimate the statistical power of a study design to detect systematic within-subject variability in these measures and thus to determine whether a study had a realistic chance to detect within-subject associations between fluctuations in cognitive control measures and fluctuations in other variables including situational predictors and behavioural outcomes.

Indeed, fluctuations in cognitive control have been shown to be influenced by several situational variables including concurrent stress levels (Shields et al., 2016), positive and negative affect (Dreisbach, 2006), motivation (Botvinick & Braver, 2015; Inzlicht & Berkman, 2015), mental exhaustion (Inzlicht & Berkman, 2015), arousal levels (Lo et al., 2016), acute physical activity and exercise (Ludyga et al., 2016), circadian timing (Blatter & Cajochen, 2006), sleep duration of the previous night (van Dongen et al., 2003; Lim & Dinges, 2010), substance use and intoxication (McPhee & Hendershot, 2023), and situational distraction e.g. by noise (Szalma & Hancock, 2011). Reporting of the LST-R-indices reliability, occasion specificity, and common consistency of cognitive control measures is essential to understand to which degree test scores are due to measurement error, situation-specific influences and stable person characteristics.

Similar to within-subject reliability, occasion specificity is of particular eminence for understanding how much variance of an intensive longitudinal study is due to systematic fluctuations. Previous studies that aimed to quantify occasion specificity of cognitive control, however, have shown only little variance being explained by the occasion compared to a much larger trait-component. Faßbender et al. (2023), for instance, using a laboratory setting

only found 2% of the total variance explained by occasion specificity in the Stroop task congruent-incongruent reaction time (RT) differences, while common consistency (i.e. trait-component) was estimated at 50%. For the Go/No-Go task error rate, 6% of the variance was explained by occasion specificity, while the common consistency explained 55%. Similar estimates were found by Meyerhöfer et al. (2016). However, an important limitation of these two studies is that they measured cognitive control in highly standardized laboratory settings to minimize occasion specific influences. For instance, they only used three measurement timepoints in total which were all scheduled at the same time of day and without occasion-specific experimental manipulations. While such procedures are adequate to measure peak performance and increase the amount of variance explained by the trait-component and in consequence lead to higher retest-reliability, laboratory settings naturally lack ecological validity. In contrast, cognitive control tasks embedded in EMA may provide higher levels of ecological validity (Zech et al., 2023). Moreover, a more even distribution of occasion specific variance and variance due to common consistency may be expected if the number of measurement occasions was similar to the number of participants in the study.

One EMA study that has investigated within-subject reliability for other cognitive tasks including working memory tasks (i.e., N-back and dot memory task) and processing speed (i.e., symbol search task) has shown moderate levels of within-subject reliability (Sliwinski et al., 2018). Sliwinski and colleagues also examined whether within-subject reliability changed across the duration of their EMA study as it is possible that with increasing duration of the EMA study participants' motivation decreases, which might negatively impact reliability. On the other hand, it is also possible that reliability is lower in the early phase of the study when participants are not yet as familiar with the tasks as later on. Importantly, within-subject reliability of working memory and processing speed was not

associated with the duration of the EMA protocol, which was shown by a constant level of within-subject reliability across 14 study days.

The present study

The present studies had the following four goals.

First, we aimed to establish within-subject reliability alongside the core indices of LST-R including measurement reliability, occasion specificity, and common consistency of two commonly used cognitive control/inhibition tasks, the Stroop and the Go/No-Go task, in the EMA context. Our specific interest hereby, apart from within-subject reliability, is focused on occasion specificity. The within-subject reliability as well as LST-R indices were estimated by dynamic structural equation models (Asparouhov et al., 2018; Castro-Alvarez et al., 2022; Geiser et al., 2013) which allowed us to account for effects of inertia by modelling auto-regressions between adjacent measurement occasions. Moreover, endorsing the possibility of trait change across the duration of the EMA study as envisaged by the revision of LST Theory (Steyer et al., 2015) we additionally modelled a linear trend of the trait variable as a fixed effect. It is conceivable that the trait-levels of the task performances may change over the course of the intensive longitudinal study, which may happen due to learning experience (which would result in an improvement in task performance over time) or fatigue and/or boredom (which would result in a decrease in task performance over time).

Second, we aimed to estimate the number of test trials of the Stroop and Go/No-Go task necessary to achieve an acceptable degree of within-subject reliability. Following Sliwinski et al., (2018) we also aimed to establish whether within-subject reliability changed across the duration of the EMA protocol.

Third, we aimed to estimate convergent validity of the Stroop and Go/No-Go task while disentangling between- (trait) and within-subject (state) level. For this we calculated between- and within-subject Pearson correlations. To estimate the convergent validity, the

correlations between the Stroop and Go/No-Go tasks were used, which both measure the ability to inhibit a dominant response tendency (Miyake et al., 2000). We hypothesized to find positive correlations particularly on the within-subject level between the Stroop and Go/No-Go task ERs, which conceptually best represent the (un-)successful inhibition of a dominant response (whereas e.g. the Go/No-Go task RT rather reflects processing speed as only RTs in Go-trials are considered, which is conceptually more closely related to Stroop RTs in congruent trials). Furthermore, the hypothesis is focused on the within-subject association for which the detection is facilitated due to higher statistical power. Previous research has only showed cross-sectional associations between Stroop and Go/No-Go indices (Lamm et al., 2006; Morooka et al., 2012) and particularly the Stroop RT was not correlated with the Go/No-Go ER on the between subject level in one study with repeated measures (Faßbender et al., 2023). Furthermore, the two cognitive control tasks were correlated with self-report measures of executive functions, self-control, and goal pursuit, which were also measured in EMA and were expected to be related with cognitive control tasks to a more modest degree (Paap et al., 2020).

Fourth, we examined how much of the within-subject variance of cognitive control tasks was explained by common contextual variables in the EMA context including time of day, location (being indoors vs. outdoors), social context (being together with other people vs. alone), and self-reported tiredness. To reach these four goals we conducted two EMA studies employing an iterative approach. We first conducted a Pilot Study to inform adaptations to the procedure of a larger Main Study.

Method

Procedure

We conducted a Pilot Study and a Main Study in accordance with the Declaration of Helsinki and with approval of the Ethics Committee of Bielefeld University (reference

number 2023-160). Initially, participants were informed about the procedure and conditions of participation as well as their rights. All participants provided informed consent and confirmed that they fulfilled prerequisite criteria. German native speakers without dyschromatopsia, aged 18 to 30 years, and who had a smartphone with an Android operating system available for the duration of the study were eligible to participate in the research.

The data collection lasted 15 days (one day baseline assessment, 14 days EMA) and took place in July 2023 (Pilot Study) and August/September 2023 (Main Study). On the first day, the participants completed a baseline questionnaire. On the second day, EMA questionnaires and cognitive tasks (Stroop, Go/No-Go) started. The EMA questionnaires were presented in movisensXS (version 1.5.23; library version 8399; movisens GmbH, Karlsruhe, Germany). The cognitive tasks were presented using Presentation for Android (version 3.0.5, Neurobehavioral Systems, Inc., Berkeley, CA, USA), but were initiated within the experience sampling questionnaires. Participants received five prompts per day in the Pilot Study and four prompts per day in the Main Study to complete a short questionnaire and the cognitive tasks for 14 consecutive days. Each prompt had the following sequence: (1) completing the questionnaires, (2) performing the Stroop task (two blocks), and (3) performing the Go/No-Go task (two blocks). In the Main Study, participants additionally performed the Fruit Tapping Task (a gamified version of the Stop-Signal-Task; Smittenaar et al., 2014; Zech et al., 2023), which is not part of the current study.

In the Pilot Study, participants were able to manually start the first questionnaire of each day and were asked to do so as soon as possible after waking up. In case they did not start the first questionnaire manually, participants received a reminder prompt between 8.30 a.m. and 9.30 a.m. For the next four questionnaires during the day, a prompt was sent at random times within the following time intervals: (1) 10.30 a.m. – 11.30 a.m., (2) 2.00 p.m. – 3.00 p.m., (3) 5.30 p.m. – 6.30 p.m., and (4) 8.30 p.m. – 9.30 p.m. In the Main Study,

participants received prompts for the four questionnaires of a day at random times within the following time intervals: (1) 9.30 a.m. – 10.30 a.m., (2) 1.00 p.m. – 2.00 p.m., (3) 4.30 p.m. – 5.30 p.m., and (4) 8.00 p.m. – 9.00 p.m.

In both studies, participants could respond to each questionnaire within 30 minutes after receiving each prompt. If participants did not respond within 30 minutes, this questionnaire was marked as missing. Participants were instructed to ignore prompts in situations that could endanger themselves or others (e.g., while driving).

Participants

Participants were recruited via email lists for study advertisements, the study participation management system of the Department of Psychology at Bielefeld University, and word-of-mouth advertisements among the students.

In the Pilot Study, the sample consisted of 21 participants (66.7% females, 33.3% males; mean age = 25.1 years, *SD* = 3.4) providing 986 assessments (mean per person = 47.0, *SD* = 21.3) after exclusion criteria were applied (see Data Preprocessing). Participants who completed at least 80% of all assessments received a 60€ voucher. Participants who completed at least 50% but less than 80% of all assessments received a 30€ voucher. Three 100€ vouchers were raffled among all participants qualifying for compensation.

In the Main Study, the sample consisted of 70 participants (68.6% females, 25.7% males, 5.7% preferred not to answer; mean age = 24.7 years, *SD* = 3.4) providing 2481 assessments (mean per person = 35.4, *SD* = 18.1) after exclusion criteria were applied (see Data Preprocessing). Participants who completed at least 50% of all assessments were eligible for receiving a compensation. The amount of compensation depended on the number of completed assessments, with 2.50 € per assessment.

Measures and instruments

Color Stroop Task

Each trial of the Stroop task started with a central fixation cross presented for a random duration between 350 and 650 ms on black background, followed by a stimulus presentation of the German word for red ("red"), blue ("blau"), green ("grün"), and yellow ("gelb"). The words were either printed in red (RGB 255, 0, 0), blue (RGB 0, 0, 255), green (RGB 0, 255, 0), or yellow (RGB 255, 255, 0). In each trial, the stimulus was presented until a response was given or for a maximum of 2500 ms in the Pilot Study and 1200 ms in the Main Study. A response could be given via four buttons placed in the corners of the smartphone display. Each button had a grey background, and the corresponding response option was in white font (top left: "red", top right: "blau", bottom left: "grün", bottom right: "gelb"). Participants were instructed to tap the button that refers to the same color as the stimulus word was printed in and not to the meaning of the word. In each session, participants were instructed to hold the smartphone horizontally with both hands so that they could use both thumbs to respond. Participants were asked to respond as quickly and accurately as possible. The stimuli could be categorized as congruent (e.g., "rot" displayed in red font color) or incongruent (e.g., "rot" displayed in green font color). In the Pilot Study, the Stroop task consisted of two blocks of 12 congruent trials (3x each congruent combination) and 12 incongruent trials (1x each incongruent combination) in randomized order (i.e., 48 trials in total). In the Main Study, the task of two blocks of 24 congruent trials (6x each congruent combination) and 24 incongruent trials (1x each incongruent combination) in randomized order (i.e., 96 trials in total). In both studies, each block started with four additional congruent practice trials (fixed order: red, blue, green, yellow) so that the participants were reminded of the button positions and could memorize them. These four

practice trials were not considered for analysis. A schematic of the Stroop task is displayed in Figure 1.

Figure 1###

Go/No-Go Task

In each session, participants were first instructed to hold the smartphone horizontally with both hands and to use the thumb of their dominant hand to respond. Participants were asked to tap the screen (any position on the screen) when a red circle (i.e., go stimulus; RGB 255, 0, 0) appeared and contain their response when a blue circle (i.e., no-go stimulus; RGB 0, 0, 255) appeared. In the Pilot Study, the task consisted of two blocks of 28 go stimuli and 12 no-go stimuli in randomized order (i.e., 80 trials in total and 30% of no-go stimuli). In the Main Study, the task consisted of two blocks of 80 go stimuli and 20 no-go stimuli in randomized order (i.e., 200 trials in total and 20% of no-go stimuli). The go stimuli were presented more frequently than the no-go stimuli, so that a tendency to respond to go stimuli was elicited. Each trial consisted of a central white fixation cross presented for a random duration between 350 and 650 ms followed by a stimulus presentation until a response was given or for a maximum of 850 ms in the Pilot Study and 650 ms in the Main Study. A schema of the Go/No-Go task is displayed in Figure 1.

Executive function problems

To assess executive function problems, an adapted version of the Webexec (Buchanan et al., 2010) was used. The six items of the Webexec were adapted so that participants were asked in each questionnaire about problems in the last 30 minutes in different aspects of executive function (keeping attention on a particular task; concentrating on a task; carrying out more than one task at a time; losing one's train of thought; seeing tasks through to completion; controlling impulsivity). Participants could answer each item on a visual

analogue scale ranging from 0 (i.e., no problems experienced) to 100 (i.e., severe problems experienced). An overall score was calculated by summing the scores of all items, which showed excellent between-subject reliability ($\Omega_{between}$ = .98) and within-subject reliability (Ω_{within} = .90) in the Pilot Study and excellent between-subject reliability ($\Omega_{between}$ = .97) and within-subject reliability (Ω_{within} = .90) in the Main Study.

Self-reported goal pursuit and self-control measures

Participants were asked whether they had been busy pursuing a relevant goal in the last 30 minutes ("Have you been busy pursuing a relevant goal in the last 30 minutes?"). If yes, participants could also indicate how much progress they had made towards their goal ("In the last 30 minutes I have achieved this goal or made good progress towards it.") on a visual analogue scale ranging from 0 (i.e., no progress) to 100 (i.e., very great progress).

Based on items that were adapted from Hofmann et al. (2012a, 2012b), participants were also asked whether they were currently experiencing a desire or had experienced a desire in the last 30 minutes ("Are you currently experiencing a desire or have you experienced a desire in the last 30 minutes?"). If so, participants were also asked how strong this desire is/was ("How strong is/was this desire?"), to what extent it conflicts with current goals ("To what extent does this desire conflict with your current goals?"), and how well they were able to resist the desire ("In the last 30 minutes, I was able to resist this desire well."). Participants could respond to each item on a visual analogue scale from 0 (i.e., not at all) to 100 (i.e., very much).

Based on items validated by Wolff et al. (2022) that were adapted to the ecological momentary assessment context, participants were asked how self-disciplined they were in the last 30 minutes ("In the last 30 minutes, I was self-disciplined.") and to what extent they exerted willpower to achieve their goal ("In the last 30 minutes, I exerted willpower to stay focused on my goals."). Participants could respond to each item on a visual analogue scale

from 0 (i.e., not at all) to 100 (i.e., very much). A sum score of both items was computed for further analysis, which showed good between-subject reliability ($\Omega_{between}$ = .89) and good within-subject reliability (Ω_{within} = .80) in the Pilot Study and good between-subject reliability ($\Omega_{between}$ = .79) and good within-subject reliability (Ω_{within} = .79) in the Main Study. To also assess a situational strategy of self-control, which according to the process model of self-control (Duckworth et al., 2016) plays an important role in preventing temptations to arise and decrease the need to exert cognitive control related to suppression of thoughts or emotions, participants were also asked to what extent they made sure that their environment was conducive to achieving their goals ("I made sure that my environment was conducive to focused work (or to achieving my goals)."). Participants could respond to each item on a visual analogue scale from 0 (i.e., not at all) to 100 (i.e., very much).

Contextual variables

The current social context was assessed by asking the participants to indicate who they are currently with ("Who are you currently with?") with multiple options. For further analysis, responses were classified into "being alone" and "being with others."

The current location was assessed by asking the participants where they were at ("What is your current location?") with multiple options. For further analysis, responses were classified into "being inside" and "being outside".

Tiredness was assessed by asking the participants to what extent they currently feel tired ("How tired do you feel at the moment?"). Participants responded on a visual analogue scale ranging from 0 (i.e., not at all) to 100 (i.e., very much).

Data preprocessing

Data preprocessing and all statistical analyses were performed using R (R Core Team, 2023) and Mplus (Muthén & Muthén, 1998-2017).

Task scoring

When scoring the Stroop and Go/No-Go task, we focused on indices that were frequently used in the literature for the two cognitive control tasks (see Faßbender et al., 2023). For the Stroop task, we calculated the error rate in incongruent trials (hereafter referred to as Stroop ER) and the reaction time difference between congruent and incongruent trials (hereafter referred to as Stroop task RT-Diff). For the Go/No-Go task, we calculated the error rate in No-Go trials (hereafter referred to as Go/No-Go ER) and the reaction time in Go trials (hereafter referred to as Go/No-Go RT). Before calculating the scores, we applied restrictions following Faßbender et al. (2023): (1) Trials with a RT < 150 ms were marked as invalid, (2) RT scores were only based on correct responses, and (3) at least ten trials per block and condition (i.e., "congruent" and "incongruent" in the Stroop task and "go" and "no-go" in the Go/No-Go task) had to be available to be considered for scoring. In order to allow for decomposition of within-subject and between-subject variance, both Stroop and Go/No-Go tasks were measured in two separate blocks within each measurement occasion and from each block one indicator/aggregate was derived.

Exclusion criteria

We applied several exclusion approaches to account for careless responding that could bias the statistical analyses: (1) For each prompt, we calculated the time the participants needed to complete the questionnaire which was filled out before completing the tasks. If the average processing time per item was < 1s, the respective prompt was excluded from further analyses (Jaso et al., 2022). (2) For the Stroop task, we calculated an index indicating the maximum sequence of subsequent identical responses per block using R-package *careless* (Yentes & Wilhelm, 2023). Prompts with an index of more than three standard deviations above the mean were also excluded from further analysis. (3) If the error rate in one of the two tasks was three standard deviations above the mean, the scores for the

respective task were excluded from further analyses. Based on these criteria, a total of 7 occasions of the Stroop task and 8 occasions of the Go/No-Go task were excluded from the pilot study. In the main study, a total of 54 occasions of the Stroop task and 55 occasions of the Go/No-Go task were excluded.

Statistical analysis

(Overall) Reliability, common consistency, occasion specificity, and within-subject reliability

Dynamic structural equation modelling. To establish reliability, occasion specificity, and common consistency while accounting for trait and state changes over time, we computed AR(1) DSEM models in Mplus (Asparouhov et al., 2018; Muthén & Muthén, 1998-2017). The data and analysis code can be found on Open Science Framework (https://osf.io/rjbx2/). Only participants with at least five complete assessments were considered. Separate but structurally comparable models were computed for the ER in incongruent trials and the RT-Diff between congruent and incongruent trials of the Stroop task and the ER and RT of the Go/No-Go task. Each outcome is represented by two observed variables, each aggregating the scores (ER, RT, RT-Diff) for the first or second block, respectively. In all models, variance of the i*th* observed variable measured at occasion t of person n (Y_{itn}) is decomposed into a within-person component (within-level) and a between-person component (between-level). On the within-level, the observed variables Y_{w1tn} and Y_{w2tn} are further decomposed into shared occasion factors OCC_{tn}, the linear effect of *time* with regression weight β, and the measurement error ϵ_{itn}. There are autoregressive effects between OCC_{tn} variables with fixed slope across persons (φ). The OCC_{tn} variables comprise the state residual variables as defined in LST-R theory but are not identical. In fact, the residuals of the autoregression correspond to the state residuals as defined in LST-R theory (see Eid et al., 2017; Stadtbaeumer et al., 2024; for details). On the between-level, the

outcomes Y_{b1tn} and Y_{b2tn}, measure ξ_{1n}, the person-specific trait variable at *t*=1. The model allows for changes in traits, which means that outcomes Y_{b1tn} and Y_{b2tn} for $t = 2, \ldots, T$ may measure ξ_{1n} plus trait changes due to the linear time trend and autoregression. The resulting model equations are as follows:

Within-level

$$Y_{w1tn} = 1 * OCC_{tn} + \beta * time_t + \epsilon_{1tn}$$

$$Y_{w2tn} = 1 * OCC_{tn} + \beta * time_t + \epsilon_{2tn}$$

$$OCC_{tn} = \varphi * OCC_{t-1,n} + \zeta_{tn} \text{ for } t = 2, \ldots, T$$

Between-level

$$Y_{b1tn} = 1 * \xi_{1n}$$

$$Y_{b2tn} = 1 * \xi_{1n}$$

We assume equal variances for ζ_{tn}-variables and ϵ_{itn}-variables. The mean of ξ_{1n} is freely estimated. The combined general equations correspond to:

$$Y_{1tn} = Y_{b1tn} + Y_{w1tn} = \xi_{1n} + OCC_{tn} + \beta * time_t + \epsilon_{1tn}$$

$$Y_{2tn} = Y_{b2tn} + Y_{w2tn} = \xi_{1n} + OCC_{tn} + \beta * time_t + \epsilon_{1tn}$$

Considering the autoregressive structure with $OCC_{1n} = \zeta_{1n}$ for the first time point and $OCC_{tn} = \varphi * OCC_{t-1,n} + \zeta_{tn}$ for subsequent time points, this results in the full model equation for $t = 1,\ldots,T$ time points for any of the outcomes:

$$Y_{itn} = \xi_{1n} + \beta * time_t + \sum_{j=1}^{t} \varphi^{j-1} * \zeta_{jn} + \epsilon_{itn}$$

Figure 2 illustrates the model conceptually.

Due to the autoregressive structure, the variance in the observed variable, Var(Y_{itn}), comprises variance in ξ_{1n}, the residual variance of ϵ_{itn}, and occasion-specific variance stemming from the state-residuals. Assuming a stable autoregressive effect, we approximated Var(Y_{itn}) by incorporating variance components from the previous four measurement occasions:

$$Var(Y_{itn}) = \text{Var}(\xi_{1n}) + \text{Var}(OCC_{tn}) + \text{Var}(\epsilon_{itn}) \text{ with}$$

$$Var(OCC_{tn}) \approx \varphi^8 * \text{Var}(\zeta_{t-4,n}) + \varphi^6 * \text{Var}(\zeta_{t-3,n}) + \varphi^4 * \text{Var}(\zeta_{t-2,n}) + \varphi^2 * \text{Var}(\zeta_{t-1,n}) \pm \text{Var}(\zeta_{tn})$$

The within-person reliability is then defined as the proportion of systematic within-subject variance. Due to the autoregressive effects, this includes state residual variance and variance components from the previous four measurements:

$$Rel_{within}(Y_{witn}) = \frac{Var(OCC_{tn})}{\text{Var}(Y_{witn})}$$

The within-person reliability hence measures how reliably both indicators Y_{1tn} and Y_{2tn} capture fluctuations in responses within-persons across time. In comparison, (overall) reliability, occasion specificity, and consistency relate not only to the within-level variance, but the total variance $Var(Y_{itn})$ and are calculated as follows:

$$Rel(Y_{itn}) = \frac{Var(\xi_{1n}) + Var(OCC_{tn})}{Var(Y_{itn})}$$

$$Spe(Y_{itn}) = \frac{Var(\zeta_{tn})}{Var(Y_{itn})}$$

$$Con(Y_{itn}) = \frac{Var(\xi_{tn})}{Var(Y_{itn})}$$

Given a specific time point *t*, the linear time trend does not add to the variance of Y_{itn} because it is constant. By definition, it holds that $Rel(Y_{itn}) = Spe(Y_{itn}) + Con(Y_{itn})$. Note that occasion specificity, in contrast to the within-person reliability, quantifies the proportion of state residual variance in the total variance. This excludes the autoregressive effects and hence measures the influence of purely situational effects and person-situation interactions.

Association of within-subject reliability with test length (number of Stroop and Go/No-Go trials) and stability of within-subject reliability across the duration of the study

To investigate how many trials are necessary to reach adequate levels of measurement reliability (i.e., agreement between the two blocks of each task), we first calculated the relevant task scores considering increasing numbers of trials per block. For example, we calculated the Stroop ER and RT-Diff when two, four, six, ..., and *n* trials (always 50% congruent and 50% incongruent trials) of a block were considered. For each number of trials, the within-subject reliability R_c (i.e., reliability of change; similar to Sliwinski et al., 2018) were computed using the mlr-function of R-package *psych* (Revelle, 2023) for the two blocks of the same cognitive task that were measured during the same measurement occasion:

$$R_c = \frac{\sigma_{p*t}^2}{\sigma_{p*t}^2 + \frac{\sigma_\epsilon^2}{m}}$$

σ_{p*t}^2 and σ_ϵ^2 represent the variance components of the person by measurement occasion and the measurement error, respectively. The number of items (i.e., the two block scores) is denoted as m.

Also following Sliwinski et al. (2018), we examined the stability of the within-subject reliability across the ambulatory assessment period by estimating the within-subject reliability separately for each day of assessment.

Convergent validity

To test convergent validity, between- and within-subject Pearson correlations were computed using the statsBy-function of R-package *psych* (Revelle, 2023). Correlations between the scores of the Stroop and the Go/No-Go task and self-report measures were examined. The self-report measures were executive function problems (i.e., Webexec score), progress in goal pursuit, desire strength, conflict of a desire with one's goals, resistance to a

desire, conduciveness of the environment (i.e., involving situational strategies of self-control), and the sum score of the two items measuring self-discipline and willpower.

Total and within-subject variance explained by contextual characteristics

Finally, to investigate how much of the within-subject variance is explained by contextual variables, multilevel models were computed using R-package *lme4* (Bates et al., 2015). In separate models, the test scores were included as outcome variables. In each model, time of day (i.e., beep number), social situation, location, and tiredness were included as predictors. For this analysis, all continuous variables (task scores and tiredness) were first within-subject centered and then grand-mean standardized. For each model, we report the explained within-subject variance for the fixed effects based on Nakagawa et al. (2017).

Results

Descriptive statistics

Demographics and descriptive statistics are displayed in Table 1. The distribution of task scores is displayed in Figure 3.

Table 1

Figure 3

In Figure 4, averaged task measures are displayed across the whole study period. The time trend over the 14 days of measurement indicate a learning curve for Stroop task RT-Diff with RT-Diff decreasing by about 50% across the 14 days in the Pilot Study, while at the same time no increase in Stroop ERs took place. In the Main Study, the decrease of RT-Diff appeared to be much smaller. Go/No-Go RTs and ERs increased slightly across the 14 days in the Main Study.

Figure 4

(Overall) Reliability, common consistency, occasion specificity, and within-subject reliability

All parameters of interest resulting from the DSEM models for both studies are displayed in Table 2. In the Pilot study and the Main study, respectively, the Stroop ER total variance included 50.3% and 62.7% of systematic variance (i.e., indicating a reliability of .503 and .627), which can be dissected into 39.0% and 51.3% explained by common consistency (i.e., based on trait-level variance) and 10.8% and 11.1% explained by occasion specificity (i.e., based on state residual variance), respectively (all indices on scale level, that is both blocks of the respective tasks combined). The within-subject reliability was .192 and .297, respectively. The Stroop task RT-Diff total variance included 54.6% and 40.7% of systematic variance (i.e., indicating a reliability of .546 and .407), 48.2% and 37.6% were accounted for by common consistency and 2.3% and 1.4% by occasion specificity in the Pilot study and the Main study, respectively. The within-subject reliability was .123 and .081, respectively.

The Go/No-Go ER total variance included 69.0% and 84.6% of systematic variance (i.e., indicating a reliability of .690 and .846), 39.3% and 68.5% were accounted for by common consistency and 29.4% and 16.1% by occasion specificity in the Pilot study and the Main study, respectively. The within-subject reliability was .534 and .576, respectively. The Go/No-Go RT total variance included 86.3% and 86.9% of systematic variance (i.e., indicating a reliability of .863 and .869), 55.5% and 49.7% were accounted for by common consistency and 30.7% and 37.2% by occasion specificity in the Pilot study and the Main study, respectively. The within-subject reliability was .712 and .763, respectively.

Table 2

Association of within-subject reliability with test length (number of Stroop and Go/No-Go trials) per measurement occasion and stability of within-subject reliability across the duration of the study

Reliability-Task-Length-Plots are shown in Figure 5 and indicate that the within-subject reliability increased with the number of trials. While for the Stroop Task moderate levels of within-subject reliability were never reached, for the Go/No-Go ER moderate levels of within-subject reliability were reached after 11 No-Go trials in the Pilot study and 15 No-Go trials in the Main study while for the Go/No-Go RT that level was reached after 7 Go trials in the Pilot and 23 Go trials in the Main study. The settings of the Go/No-Go Task in the Pilot Study (i.e. 30% of No-Go trials in the Pilot study vs. 20% of No-Go trials in the Main study) appeared to have facilitated obtaining acceptable within-subject reliability within a shorter test duration considering the overall Go/No-Go Task duration of 1.8 minutes in the Pilot study vs. 4.3 minutes in the main study.

Figure 5

Figure 6 shows the within-subject reliability estimated separately for each of the 14 days of the EMA study. Across the 14 study-days within-subject reliability did not show systematic change that was consistent between pilot and main study for any of the indicators. Particularly in the main study the within-subject reliability of the Go/No-Go Task ER and RT was on a consistently high level across the 14 days. On average the within-subject reliability estimated separately for each of the 14 was lower than the overall average within-subject reliability, which is attributed to a lower level of systematic within-subject variability within a given day as compared to systematic within-subject variability within and across days.

Figure 6

Convergent validity

Within-subject level

Concerning convergent validity and as hypothesized, the Stroop ER was correlated with the Go/No-Go ER on the within-subject level in the Pilot Study ($r = .08$, $p < .05$) and in the Main Study ($r = .12$, $p < .001$). In the Main Study only, the Stroop ER was also correlated with the Go/No-Go RT ($r = .07$, $p < .001$) and the Stroop task RT-Diff was correlated with the Go/No-Go ER ($r = –.07$, $p < .01$).

Concerning the self-report measures, the Stroop ER ($r = –.06$, $p < .05$) and Stroop task RT-Diff ($r = –.10$, $p < .01$) were correlated with the Webexec score in the Pilot Study. In the Main Study, the Stroop task RT-Diff was also correlated with the Self-Discipline/Willpower score ($r = –.08$, $p < .05$) in the Main study. No other significant correlations between the task scores and self-report measures were found (Figure 7).

Between-subject level

Concerning convergent validity, the Stroop ER was correlated with the Go/No-Go ER on the between-subject level in the Pilot Study ($r = .79$, $p < .001$) and in the Main Study Study ($r = .57$, $p < .001$). Also, the Stroop task RT-Diff was correlated with the Go/No-Go ER ($r = .53$, $p < .05$) in the Pilot Study.

Concerning the self-report measures, the Go/No-Go ER was correlated with the Webexec score on the between-subject level in the Pilot Study ($r = .54$, $p < .05$), but not in the Main Study. In the Main Study, however, the Go/No-Go ER was correlated with the progress in achieving one's goals ($r = –.30$, $p < .01$) and the Go/No-Go RT was correlated with the Webexec score ($r = –.29$, $p < .05$).

Figure 7

Variance explained by contextual characteristics

The contextual factors explained 6.0% and 3.4% of the Stroop ER variance in the Pilot and Main Study, respectively. Feeling more tired than usual ($\beta = -0.05$, $p < .05$) and being around other people ($\beta = -0.13$, $p < .01$) were associated with a lower Stroop ER in the Main Study.

The contextual factors explained 2.7% and 2.2% of the Stroop RT-Diff variance in the Pilot and Main Study, respectively. None of the contextual factors was significantly associated with Stroop RT-Diff in both studies.

The contextual factors explained 6.4% and 4.0% of the Go/No-Go ER variance in the Pilot and Main Study, respectively. In the Main Study, the Go/No-Go ER differed between different times of the day. The Go/No-Go ER was significantly higher at beep 3 ($\beta = 0.20$, $p < .01$) and beep 4 ($\beta = 0.14$, $p < .05$) in contrast to beep 1.

The contextual factors explained 13.9% and 5.2% of the Go/No-Go RT variance in the Pilot and Main Study, respectively. In the Pilot Study, we found differences in the Go/No-Go RT for different times of the day. The Go/No-Go RT was significantly lower at beep 3 (i.e. in the late afternoon; $\beta = -.24$, $p < .05$) in contrast to beep 1 (i.e. in the morning).

Discussion

(Overall) Reliability, common consistency, occasion specificity, and within-subject reliability

To our knowledge, this study is the first to examine within-subject reliability, occasion specificity, and validity of fluctuations of the Stroop task and the Go/No-Go task in an EMA context with multiple daily assessments over several days and applying LST-R models. Our findings indicate moderate overall reliability, but poor within-subject reliability estimates for the Stroop task ER and Stroop task RT-Diff in both studies. For the Go/No-Go

task ER and RTs overall reliability was excellent and within-subject reliability was moderate to good in both studies. Thus, our findings indicate a moderately good ability of the Go/No-Go task when administered via EMA to reliably capture within-subject fluctuation of cognitive control. By contrast, the Stroop task as administered as in the present studies was not able to reliably capture moment to moment fluctuation of cognitive control. Instead its moderate (overall) reliability was mainly driven by common consistency (i.e., trait-level variance).

Compared to a previous study estimating (overall) reliability, common consistency, and occasion specificity (Faßbender et al., 2023), it sticks out that the amount of variance explained by occasion specificity was much larger and overall quite substantial in the current two studies at least regarding the Go/No-Go task. This pattern did not show up for the Stroop task, where occasion specificity of the Stroop task ER and RT-Diffs were on a very low level and similar to Faßbender et al.'s study. As the most important difference between Faßbender et al. (2023) and the current study, they studied state-level fluctuations only over three measurement time points using laboratory-based assessment, while the current two studies used an EMA design with a maximum of 70 measurement occasions in the Pilot and 56 in the Main study across 2 weeks and smartphone-based assessment. Due to the increased number of measurement occasions and the more variable context and time of day of measurement it is unsurprising that the two current studies found a larger amount of variance explained by state-level fluctuations at least with regard to the Go/No-Go Task.

Association of within-subject reliability with test length (number of Stroop and Go/No-Go trials) per measurement occasion and stability of within-subject reliability across the duration of the study

Increasing the number of test trials was only associated with a slow rise of within-subject reliability of the Stroop Task ER and RT. Reliable assessment of moment to moment

fluctuations of cognitive control by further increasing the number of trials of the Stroop task appears not to be practical, given that ecological momentary assessment exposes participants to considerable levels of burden particularly when multiple daily measures are taken which run for several minutes per measurement occasion. Several observations suggest that the increased number of trials within the main study might have already decreased participants' motivation: First, the completion rate decreased substantially from the Pilot Study to the Main Study. Second, a larger number of occasions had to be excluded in the Main study as criteria for careless responding were met. Third, the average Stroop ER was substantially increased in the Main study compared to the Pilot study although the task itself had been left unchanged. These observations do also apply to the Go/No-Go task in the Main Study as compared to the Pilot Study. However, the Go/No-Go task was slightly changed (i.e., the maximal allowed response time was decreased from 850 ms to 650 ms and the ratio of No-Go-trials to Go-trials was decreased from 30% No-Go-trials to only 20% No-Go-trials), which could at least partly explain the increase of the average error rate. With a focus on optimizing within-subject reliability while at the same time keeping subject burden as low as possible the Go/No-Go task setup of the Pilot study was preferable over the setup of the Main study as acceptable levels of within-subject reliability were obtained with a smaller number of trials and in shorter time.

Within-subject reliability of the two tasks did not show a systematic increase or decrease across the 14 study days in both studies when within-subject reliability was calculated separately for each study day. Particularly in the Main study the course of the day-wise within-subject reliability of the Go/No-Go task ER and RT across the 14 study days was relatively constant and on a high level, which was similar to the findings of Sliwinski et al. (2018) for spatial working memory and processing speed.

Convergent validity

Regarding validity estimates, the only correlation that was observed consistently across the two studies concerned Stroop ER and Go/No-Go ER. As we expected, both scores were positively correlated. Given the poor within-subject reliability of the Stroop task error rate and Stroop task RT-Diff and only moderate within-subject reliability of the Go/No-Go task error rate and Go/No-Go RTs, it is unsurprising that the Stroop task and Go/No-Go task scores do not correlate strongly on the within-subject level due to attenuation. On the between-subject level, the same positive correlation was also found in both studies, although it was much stronger there. These strong between-subject correlations between Stroop ER and Go/No-Go ER in both of our studies suggest a common underlying trait accounting specifically for the ER of both tasks. By contrast, the Stroop RT was not correlated with Go/No-Go RT in our studies while previous cross-sectional research reported positive associations between the Stroop interference score and Go/No-Go RT (Lamm et al., 2006) and between Stroop RT and Go/No-Go RT (Mooroka et al., 2012). Moreover, we could not find an association between the Stroop RT and Go/No-Go ER on both the within-subject and the between-subject level in both studies, which is consistent with Faßbender et al. (2023) and the notion that the Stroop RT and Go/No-Go ER measure distinct underlying constructs. Future research will also have to establish within-subject (i.e., situation specific) convergent validity for the Go/No-Go task by correlating it with other cognitive control tasks than the Stroop task.

No associations of the Stroop task and the Go/No-Go task scores with self-report measures of executive functions and self-control were found consistently across the two studies. These findings echo results from cross-sectional research showing no association between self-reported self-control and cognitive control (Eisenberg et al., 2019; Paap et al., 2020; Saunders et al., 2018). Thus, the findings of the present study confirm the distinctness of cognitive control as measured by cognitive control tasks from self-reported self-control

and cognitive control. Recent research suggested that self-report self-control and cognitive control tasks assess different underlying processes (Saunders et al., 2018). While self-reports may capture subjective perceptions of self-control and progress on tasks, laboratory tasks may target specific cognitive mechanisms. In real-life, goal progress is facilitated through the use of strategies that decrease the extent of goal conflict preventively (Duckworth et al. 2016). Moreover, for real-life goal progress motivational processes may be more important than the level of cognitive control. If individuals are intrinsically motivated for a task, they may be less tempted by conflicting goals (Milyavskaya et al. 2015, Werner & Milyavskaya 2019; Inzlicht et al., 2021). It has to be noted, however, that associations might be stronger in samples with higher levels of cognitive control problems involving for instance clinical samples with addictive behavior. Moreover, it also has to be noted that on the between-subject level statistical power was limited particularly in the Pilot study. Between-subject correlations of around $r = -.30$ between Go/No-Go ER and goal progress and between Go/No-Go RT and self-reported executive function problems (i.e., the Webexec Score) might indicate reliable between-subject covariation between these variables which could not be confirmed in the dataset of the Pilot study due to limited statistical power.

Variance explained by contextual variables

We found only few associations of contextual variables including time of day, location (being indoors vs. outdoors), social context (being together with other people vs. alone), and self-reported tiredness with Stroop or Go/No-Go test performances that were consistent across the two studies. A result which we found consistently in both studies involved that the Stroop ER was lower for participants who felt tired. It is possible that participants who indicated to feel tired in the questionnaire, increased their efforts to compensate for tiredness. However, a priori we would have rather expected the opposite direction of the association. With regard to environment context, it is possible that more

nuanced assessment of environmental variables (e.g., noise-level, glare, and social interference by interaction partners) would facilitate finding stronger and more consistent associations. However, as environmental variables could bias the two test blocks within a measurement occasion into the same as well as into the opposite direction (e.g., an interaction partner could interfere during both test blocks of the measurement occasion or only during one), it is difficult to discern whether interference by these variables would increase or decrease the association between the two test blocks within a measurement occasion.

Limitations

One has to be aware of the following limitations of our two studies. First, both studies included predominantly female university students. It is possible that results would look differently in other populations. Second, the compliance rate of around 67% in the Pilot Study and 51% in the Main Study was substantially lower than the normally reported compliance rate of around 79% in EMA research (Wrzus & Neubauer, 2023), which might be due to the longer EMA measurement occasions which took about 6-7 minutes per occasion in total in the Pilot Study and 15-16 minutes in the Main Study. Third, the measurement of contextual variables such as noise-level, glare, and social interference by interaction partners could have been more nuanced in order to exclude their interfering effects which might be unrelated with potential psychological variables of interest such as stress-levels but might bias the test performance. Fourth, our protocol involved EMA measurement occasions, which were timed during pre-defined time windows. It is possible that with other types of EMA protocols that would be more adaptive to the psychological circumstances, different results might arise. For instance, if participants could trigger EMA assessments themselves during moments of stress, it might be possible to measure stress effects more closely after stressors were encountered. Future research will have to examine the feasibility of such designs related to EMA assessment of cognitive control. Fifth, we only analyzed the Stroop and the Go/No-

Go task. It is possible that using other cognitive control tasks the convergent validity estimates would be higher. Findings by Faßbender et al. (2023) for instance suggest that cognitive control is not a unitary construct but involves a two-factor structure with two only moderately correlated factors. In that study the Go/No-Go task was associated with the main "response inhibition" factor while the Stroop task was associated with an "interference factor". It is therefore possible that cognitive control tasks involving interfering stimuli (such as the Eriksen flanker task) would show stronger correlations with the Stroop task, while tasks without interfering stimuli (such as the Stop Signal task) would show stronger correlations with the Go/No-Go task. Sixth, a sample size of 70 may not have provided sufficient power to detect small between-subject correlations. However, the focus of this study was on the within-subject level. Seventhes, there is a potential circularity in the assessment of convergent validity, given that both the Stroop and Go/No-Go tasks were being evaluated within the same study. This could lead to circular reasoning, where the validity of each task is inferred from its correlation with the other, despite both being under examination. Eights, the within-subject reliability that was computed to investigate its association with test length and stability across the study duration differs from the reliability measures computed in DSEM. This discrepancy could lead to situations where the number of trials deemed sufficient for traditional reliability may not necessarily meet the reliability requirements as computed by DSEM. Finally, to estimate within-subject reliability and occasion specificity we analyzed the common fluctuations of two test-halves, while it would have been possible to conduct the analyses on the single trial level. Thus, future research might consider adding an additional level to the analysis by treating Stroop and Go/No-Go ERs and RTs on the trial level with trials nested in measurement occasions because it has been shown recently that this affects reliability (Rouder & Haaf, 2019; Waltmann et al., 2023; Zech et al., 2023)

Conclusion

Using an LST-R model accounting for auto-regressions and trait-change we found moderately high within-subject reliability of the Go/No-Go task ER and RT as conducted on Smartphones in the EMA context. By contrast, the Stroop task ER and RT did not reach at least borderline acceptable levels of within-subject reliability. The amount of occasion specific variance was substantially larger for the Go/No-Go task in the EMA context than in laboratory research but systematic state-level variance remains considerably smaller than systematic trait-level variance. Indicative of convergent validity the within-subject correlations between the Stroop ER and Go/No-Go ER were positive and consistent across our two studies, but effect-sizes were modest. Future research will have to test whether the findings generalize to more diverse social groups, whether findings are consistent with different cognitive control tasks, and determine the role of interfering effects of the environmental context by using a more nuanced assessment of contextual variables.

Declarations

Funding

No external funding was received for conducting this study.

Conflicts of interest

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval

Both studies were performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Bielefeld University (Reference No. 2023-160).

Consent to participate

Informed consent was obtained from all individual participants included in the study.

Consent for publication

Participants consented to the use of their anonymous data in this publication.

Data availability

The data and materials used in this article are openly available on Open Science Framework (https://osf.io/rjbx2/).

Code availability

The analysis code used in this article is openly available on Open Science Framework (https://osf.io/rjbx2/).

Authors' contributions

Justin Hachenberger: Conceptualization, Methodology, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Project administration. **Axel Mayer**: Conceptualization, Methodology, Formal analysis, Writing – Review & Editing. **Denny Kerkhoff**: Conceptualization, Methodology, Writing – Review & Editing. **Friederike Eyssel**: Conceptualization, Writing – Review & Editing. **Stefan Fries**: Conceptualization, Writing – Review & Editing. **Tina B. Lonsdorf**: Conceptualization, Writing – Review & Editing. **Hilmar Zech**: Conceptualization, Writing – Review & Editing. **Lorenz Deserno**: Conceptualization, Writing – Review & Editing. **Sakari Lemola**: Conceptualization, Methodology, Writing – Original Draft, Writing – Review & Editing, Project administration, Supervision

References

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural equation modeling*, *25*(3), 359-388.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, *67*. doi: 10.18637/jss.v067.i01

Blatter, K., & Cajochen, C. (2007). Circadian rhythms in cognitive performance: Methodological constraints, protocols, theoretical underpinnings. *Physiology & Behavior, 90*, 196–208. doi: 10.1016/j.physbeh.2006.09.009

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: From Behavior to neural mechanism. *Annual Review of Psychology*, *66*, 83–113. doi: 10.1146/annurev-psych-010814-015044

Brose, A., Schmiedek, F., Gerstorf, D., & Voelkle, M. C. (2020). The measurement of within-person affect variation. *Emotion*, 20(4), 677–699. doi:10.1037/emo0000583

Buchanan, T., Heffernan, T. M., Parrott, A. C., Ling, J., Rodgers, J., & Scholey, A. B. (2010). A short self-report measure of problems with executive function suitable for administration via the Internet. *Behavior Research Methods*, *42*, 709–714. doi: 10.3758/BRM.42.3.709

Castro-Alvarez, S., Tendeiro, J. N., Meijer, R. R., & Bringmann, L. F. (2022). Using structural equation modeling to study traits and states in intensive longitudinal data. *Psychological Methods*, *27*, 17–43. doi: 10.1037/met0000393

Dreisbach, G. (2006). How positive affect modulates cognitive control: The costs and benefits of reduced maintenance capability. *Brain and Cognition*, *60*, 11–19. doi: 10.1016/j.bandc.2005.08.003

Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2016). Situational strategies for self-control. *Perspectives on Psychological Science*, *11*, 35–55. doi: 10.1177/1745691615623247

Eid, M., Holtmann, J., Santangelo, P., & Ebner-Priemer, U. (2017). On the Definition of Latent-State-Trait Models With Autoregressive Effects: Insights From LST-R Theory. *European Journal of Psychological Assessment*, *33*(4), 285–295. doi: 10.1027/1015-5759/a000435

Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications*, *10*, 2319. doi: 10.1038/s41467-019-10301-1

Falkenstein, M., Hoormann, J., & Hohnsbein, J. (1999). ERP components in Go/Nogo tasks and their relation to inhibition. *Acta Psychologica*, *101*, 267–291. doi: 10.1016/S0001-6918(99)00008-6

Faßbender, K., Meyhöfer, I., & Ettinger, U. (2023). Latent state–trait and latent growth curve modeling of inhibitory control. *Journal of Experimental Psychology: General*, *152*, 1396–1419. doi: 10.1037/xge0001344

Geiser, C., Bishop, J., Lockhart, G., Shiffman, S., & Grenard, J. L. (2013). Analyzing latent state-trait and multiple-indicator latent growth curve models as multilevel structural equation models. *Frontiers in Psychology*, *4*. doi: 10.3389/fpsyg.2013.00975

Hofmann, W., Baumeister, R. F., Förster, G., & Vohs, K. D. (2012a). Everyday temptations: An experience sampling study of desire, conflict, and self-control. *Journal of Personality and Social Psychology*, *102*, 1318–1335. doi: 10.1037/a0026545

Hofmann, W., Vohs, K. D., & Baumeister, R. F. (2012b). What People Desire, Feel Conflicted About, and Try to Resist in Everyday Life. *Psychological Science*, *23*, 582–588. doi: 10.1177/0956797612437426

Inzlicht, M., & Berkman, E. (2015). Six Questions for the Resource Model of Control (and Some Answers): Six Questions. *Social and Personality Psychology Compass*, *9*, 511–524. doi: 10.1111/spc3.12200

Inzlicht, M., Werner, K. M., Briskin, J. L., & Roberts, B. W. (2021). Integrating Models of Self-Regulation. *Annual Review of Psychology*, *72*(1), 319–345. doi: 10.1146/annurev-psych-061020-105721

Jaso, B. A., Kraus, N. I., & Heller, A. S. (2022). Identification of careless responding in ecological momentary assessment research: From post-hoc analyses to real-time data monitoring. *Psychological Methods*, 27(6), 958–981. doi: 10.1037/met0000312

Jones, A., Tiplady, B., Houben, K., Nederkoorn, C., & Field, M. (2018). Do daily fluctuations in inhibitory control predict alcohol consumption? An ecological momentary assessment study. *Psychopharmacology*, *235*, 1487–1496. doi: 10.1007/s00213-018-4860-5

Lamm, C., Zelazo, P. D., & Lewis, M. D. (2006). Neural correlates of cognitive control in childhood and adolescence: Disentangling the contributions of age and executive function. Neuropsychologia, 44(11), 2139-2148.

Lim, J., & Dinges, D. F. (2010). A meta-analysis of the impact of short-term sleep deprivation on cognitive variables. *Psychological Bulletin*, *136*, 375–389. doi: 10.1037/a0018883

Lo, L. Y., Hung, N. L., & Lin, M. (2016). Angry versus furious: A comparison between valence and arousal in dimensional models of emotions. *The Journal of Psychology*, *150*, 949–960. doi: 10.1080/00223980.2016.1225658

Ludyga, S., Gerber, M., Brand, S., Holsboer-Trachsler, E., & Pühse, U. (2016). Acute effects of moderate aerobic exercise on specific aspects of executive function in different age

and fitness groups: A meta-analysis: Moderate exercise and executive function. *Psychophysiology*, *53*, 1611–1626. doi: 10.1111/psyp.12736

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*, 163–203. doi: 10.1037/0033-2909.109.2.163

McPhee, M. D., & Hendershot, C. S. (2023). Meta-analysis of acute alcohol effects on response inhibition. *Neuroscience & Biobehavioral Reviews*, *152*, 105274. doi: 10.1016/j.neubiorev.2023.105274

Meyhöfer, I., Bertsch, K., Esser, M., & Ettinger, U. (2016). Variance in saccadic eye movements reflects stable traits: Variance in saccadic eye movements. *Psychophysiology*, *53*, 566–578. doi: 10.1111/psyp.12592

Milyavskaya, M., Inzlicht, M., Hope, N., & Koestner, R. (2015). Saying "no" to temptation: Want-to motivation improves self-regulation by reducing temptation rather than by increasing self-control. *Journal of Personality and Social Psychology*, *109*(4), 677–693. doi: 10.1037/pspp0000045

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. Cognitive psychology, 41(1), 49-100.

Morooka, T., Ogino, T., Takeuchi, A., Hanafusa, K., Oka, M., & Ohtsuka, Y. (2012). Relationships between the color-word matching Stroop task and the Go/NoGo task: Toward multifaceted assessment of attention and inhibition abilities of children. Acta Medica Okayama, 66(5), 377-386.

Muthén, L.K. and Muthén, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, *14*(134), 20170213. doi: 10.1098/rsif.2017.0213

Nahum, M., Sinvani, R.-T., Afek, A., Ben Avraham, R., Jordan, J. T., Ben Shachar, M. S., Ben Yehuda, A., Berezin Cohen, N., Davidov, A., & Gilboa, Y. (2023). Inhibitory control and mood in relation to psychological resilience: An ecological momentary assessment study. *Scientific Reports*, *13*, 13151. doi: 10.1038/s41598-023-40242-1

Neubauer, A. B., & Schmiedek, F. (2020). Studying within-person variation and within-person couplings in intensive longitudinal data: Lessons learned and to be learned. *Gerontology*, *66*(4), 332-339.

Paap, K. R., Anders-Jefferson, R., Zimiga, B., Mason, L., & Mikulinsky, R. (2020). Interference scores have inadequate concurrent and convergent validity: Should we stop using the flanker, Simon, and spatial Stroop tasks?. *Cognitive research: principles and implications*, *5*(1), 1-27.

Plomin, R., DeFries, J. C., & Loehlin, J. C. (1977). Genotype-environment interaction and correlation in the analysis of human behavior. *Psychological Bulletin*, *84*, 309–322. doi: 10.1037/0033-2909.84.2.309

Powell, D. J. H., McMinn, D., & Allan, J. L. (2017). Does real time variability in inhibitory control drive snacking behavior? An intensive longitudinal study. *Health Psychology*, *36*, 356–364. doi: 10.1037/hea0000471

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

Revelle, W. (2023). Psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. R package version 2.3.6. .R-project.org/package=psych

Rónai, L., Hann, F., Keri, S., Ettinger, U., & Polner, B. (2023). Emotions under control? Better cognitive control predicts reduced negative emotionality but increased negative emotional reactivity within individuals [Preprint]. PsyArXiv. doi: 10.31234/osf.io/a2fdw

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. Psychonomic bulletin & review, 26(2), 452-467.

Saunders, B., Milyavskaya, M., Etz, A., Randles, D., & Inzlicht, M. (2018). Reported self-control is not meaningfully associated with inhibition-related executive function: A Bayesian analysis. *Collabra: Psychology*, *4*, 39. doi: 10.1525/collabra.134

Shields, G. S., Sazma, M. A., & Yonelinas, A. P. (2016). The effects of acute stress on core executive functions: A meta-analysis and comparison with cortisol. *Neuroscience & Biobehavioral Reviews*, *68*, 651–668. doi: 10.1016/j.neubiorev.2016.06.038

Sliwinski, M. J., Mogle, J. A., Hyun, J., Munoz, E., Smyth, J. M., & Lipton, R. B. (2018). Reliability and validity of ambulatory cognitive assessments. *Assessment*, *25*(1), 14-30.

Smittenaar, P., Rutledge, R. B., Zeidman, P., Adams, R. A., Brown, H., Lewis, G., & Dolan, R. J. (2015). Proactive and reactive response inhibition across the lifespan. *PLoS One*, 10(10), e0140383.

Stadtbaeumer, N., Kreissl, S., & Mayer, A. (2024). Comparing revised latent state–trait models including autoregressive effects. *Psychological Methods*, *29*(1), 155–168. doi: 10.1037/met0000523

Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of States and Traits—Revised. *Annual Review of Clinical Psychology*, *11*, 71–98. doi: 10.1146/annurev-clinpsy-032813-153719

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, *13*, 389–408. doi: 10.1002/(SICI)1099-0984(199909/10)13:5<389::AID-PER361>3.0.CO;2-A

Szalma, J. L., & Hancock, P. A. (2011). Noise effects on human performance: A meta-analytic synthesis. *Psychological Bulletin*, *137*, 682–707. doi: 10.1037/a0023987

Van Dongen, H. P. A., Maislin, G., Mullington, J. M., & Dinges, D. F. (2003). The cumulative cost of additional wakefulness: Dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *Sleep*, *26*, 117–126. doi: 10.1093/sleep/26.2.117

Waltmann, M., Schlagenhauf, F., & Deserno, L. (2022). Sufficient reliability of the behavioral and computational readouts of a probabilistic reversal learning task. *Behavior research methods*, 54(6), 2993-3014.

Werner, K. M., & Milyavskaya, M. (2019). Motivation and self-regulation: The role of want-to motivation in the processes underlying self-regulation and self-control. *Social and Personality Psychology Compass*, *13*(1), e12425. doi: 10.1111/spc3.12425

Wolff, W., Bieleke, M., Englert, C., Bertrams, A., Schüler, J., & Martarelli, C. S. (2022). A single item measure of self-control – validation and location in a nomological network of self-control, boredom, and if-then planning. *Social Psychological Bulletin*, *17*, e7453. doi: 10.32872/spb.7453

Wrzus, C., & Neubauer, A. B. (2023). Ecological Momentary Assessment: A Meta-Analysis on Designs, Samples, and Compliance Across Research Fields. *Assessment*, *30*, 825–846. doi: 10.1177/10731911211067538

Yentes R.D., & Wilhelm, F. (2023). careless: Procedures for computing indices of careless responding.R package version 1.2.2.

Zech, H. G., Reichert, M., Ebner-Priemer, U. W., Tost, H., Rapp, M. A., Heinz, A., Dolan, R. J., Smolka, M. N., & Deserno, L. (2022). Mobile data collection of cognitive-behavioral tasks in substance use disorders: Where are we now? *Neuropsychobiology*, *81*, 438–450. doi: 10.1159/000523697

Zech, H., Waltmann, M., Lee, Y., Reichert, M., Bedder, R. L., Rutledge, R. B., ... & ReCoDe-Consortium. (2023). Measuring self-regulation in everyday life: Reliability and validity of smartphone-based experiments in alcohol use disorder. *Behavior research methods*, *55*(8), 4329-4342.

Table 1

Demographics and descriptive statistics

	Pilot Study ($N = 21$)	Main Study ($N = 70$)
	n (%) / *M* (*SD*)	*n* (%) / *M* (*SD*)
Demographics		
Gender		
Female	14 (66.7)	48 (68.6)
Male	7 (33.3)	18 (25.7)
Not responded	0 (0.0)	4 (5.7)
Age	25.14 (3.4)	24.66 (3.4)
Ambulatory Assessment		
Average Total Duration[a] [min]	6.7 (2.4)	15.5 (4.0)
Stroop		
Complete Measurement Occasions [*n*]	47.1 (21.4)	35.6 (18.3)
Completion Rate [%]	67.4 (30.5)	50.9 (26.1)
Valid Measurement Occasions [*n*]	46.8 (21.2)	34.8 (18.3)
Excluded Measurement Occasions [*n*]	0.3 (0.8)	0.8 (2.1)
Average Task Duration [min]	2.0 (0.6)	4.0 (3.9)
ER [%]	14.6 (4.1)	19.1 (5.9)
RT-Diff [ms]	137.5 (70.0)	55.8 (35.8)
Go/No-Go		
Complete Measurement Occasions [*n*]	46.9 (21.5)	35.3 (18.4)
Completion Rate [%]	67.0 (30.8)	50.5 (26.3)
Valid Measurement Occasions [*n*]	46.5 (21.4)	34.5 (18.4)
Excluded Measurement Occasions [*n*]	0.4 (1.0)	0.8 (2.0)
Average Task Duration [min]	1.8 (0.3)	4.3 (1.8)
ER [%]	13.1 (6.3)	34.3 (15.9)
RT [ms]	319.1 (28.0)	287.6 (20.5)
Self-Report		
Webexec Score	148.3 (92.1)	152.9 (94.8)
Goal Progress	63.1 (15.8)	62.3 (18.5)
Desire Strength	62.7 (11.6)	58.9 (17.1)
Desire Resistance	49.9 (19.8)	50.0 (21.3)
Environment Adjustment	50.7 (22.1)	42.9 (22.9)
Self-Discipline/Willpower	104.7 (31.6)	117.7 (34.7)
Feeling Tired	45.2 (14.1)	47.5 (19.9)
Being outdoors [*n*]	3.0 (3.1)	3.4 (4.3)
Being alone [*n*]	23.2 (12.5)	16.6 (11.9)

Note. ER error rate; RT response time; RT-Diff difference in response times between congruent and incongruent trials. The values for the ambulatory assessment variables represent the averages mean values per subject.

[a]Including self-report, Stroop Task, Go/No-Go Task and Fruit Tapping Task.

Table 2

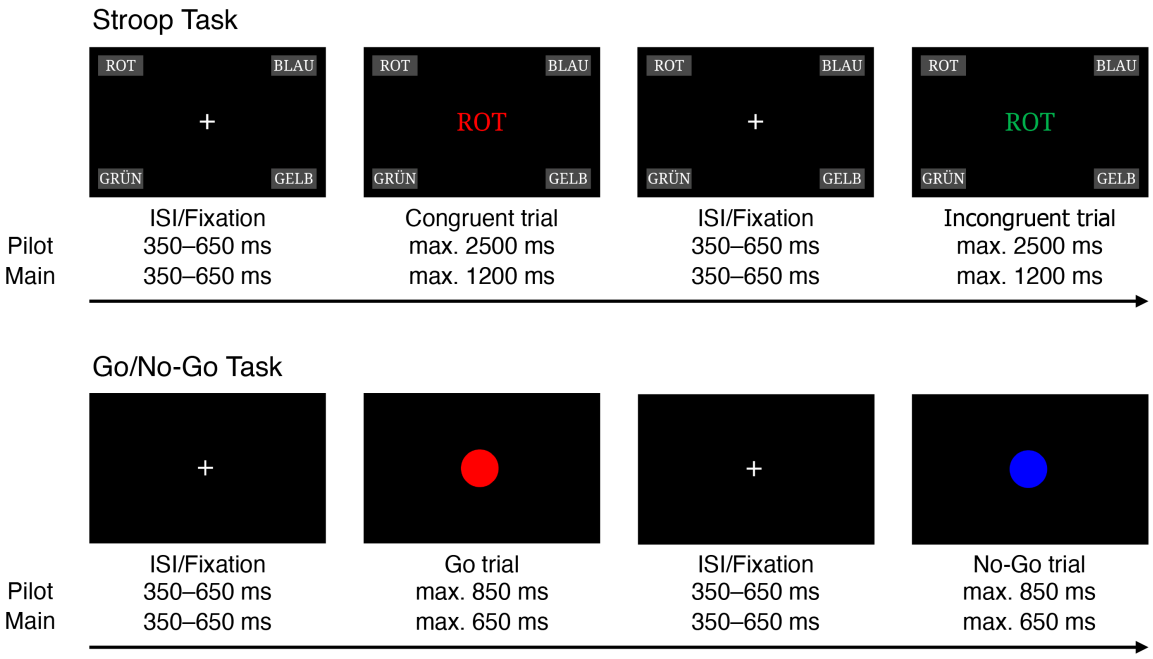
Coefficients as estimated by the dynamic structural equation model with a linear trend

Task	Score	Study	(Overall) Reliability	Common Consistency	Occasion Specificity	Within-subject Reliability
Stroop	ER	Pilot	.503	.390	.108	.192
		Main	.627	.513	.111	.315
	RT-Diff	Pilot	.546	.482	.023	.123
		Main	.407	.376	.014	.081
Go/No-Go	ER	Pilot	.690	.393	.294	.534
		Main	.846	.685	.161	.576
	RT	Pilot	.863	.555	.307	.712
		Main	.869	.497	.372	.763

Note. ER error rate; RT response time; RT-Diff difference in response times between congruent and incongruent trials.

Figure 1

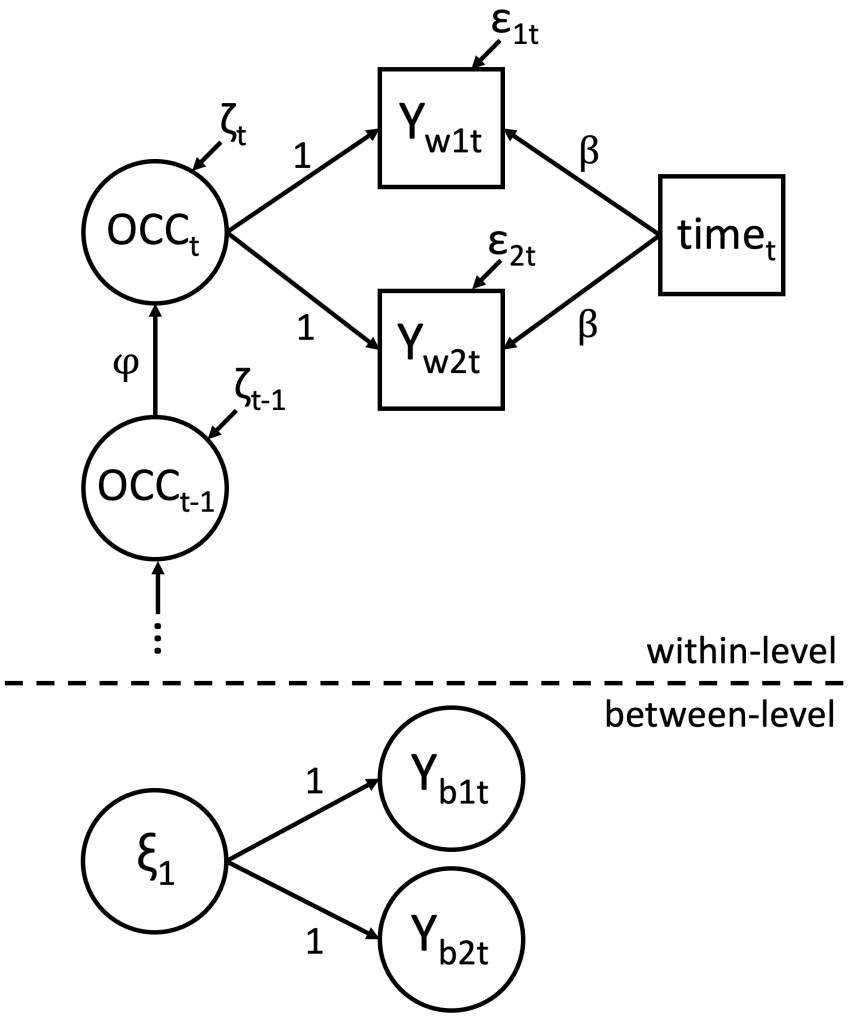
Schematic of the Stroop and Go/No-Go Task



Note. ISI interstimulus interval.

Figure 2

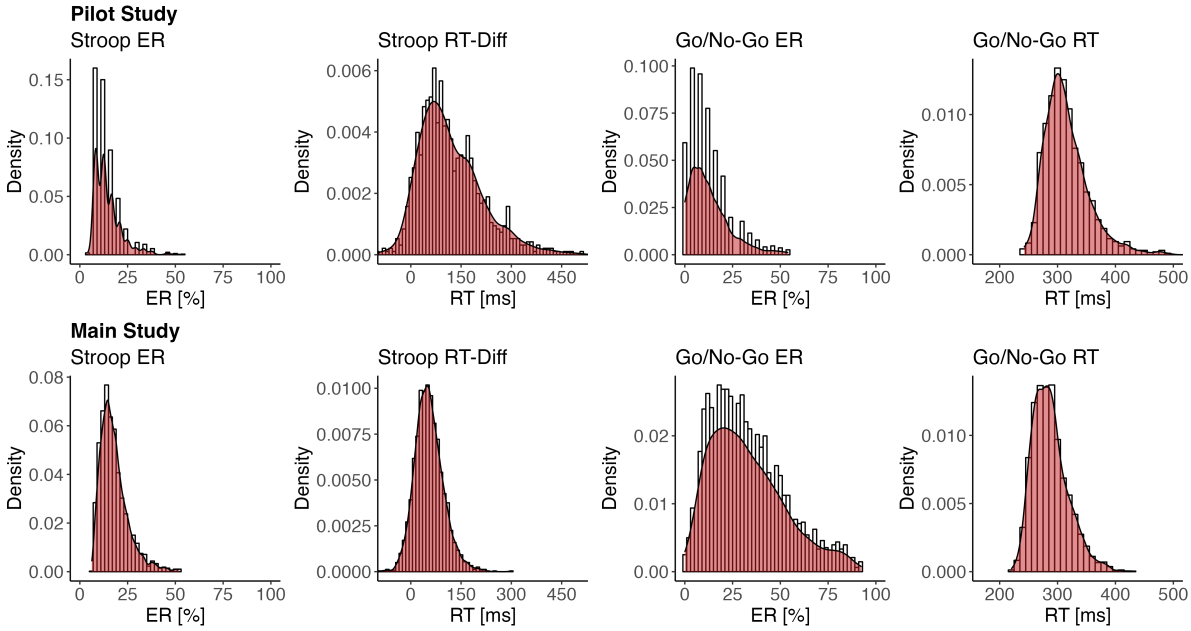
Model diagram of the dynamic structural equation models with autoregressive structure [AR(1) DSEM]



Note. The upper part represents within-person associations for a timepoint t and the previous timepoint t-1, with within-level components of the dependent variable, Y_{w1t} and Y_{w2t}, decomposed into time-specific occasion factors OCC and the effect of time. The occasion factors OCC_t and OCC_{t-1} with coefficient φ represent the autoregressive structure. The lower part represents between-person associations, with between-level components of the dependent variable, Y_{b1t} and Y_{b2t}, and person-specific trait variable at timepoint 1, ξ_1. To facilitate readability, the person-subscript n is omitted.

Figure 3

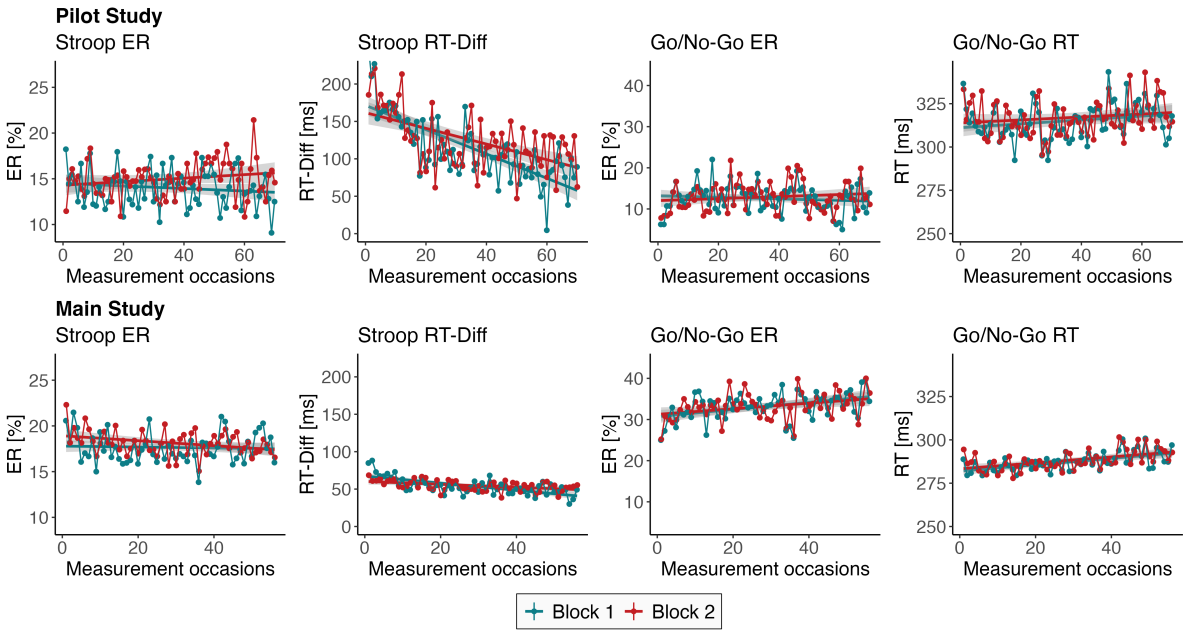
Distribution of task scores



Note. ER error rate; RT response time; RT-Diff difference in response times between congruent and incongruent trials.

Figure 4

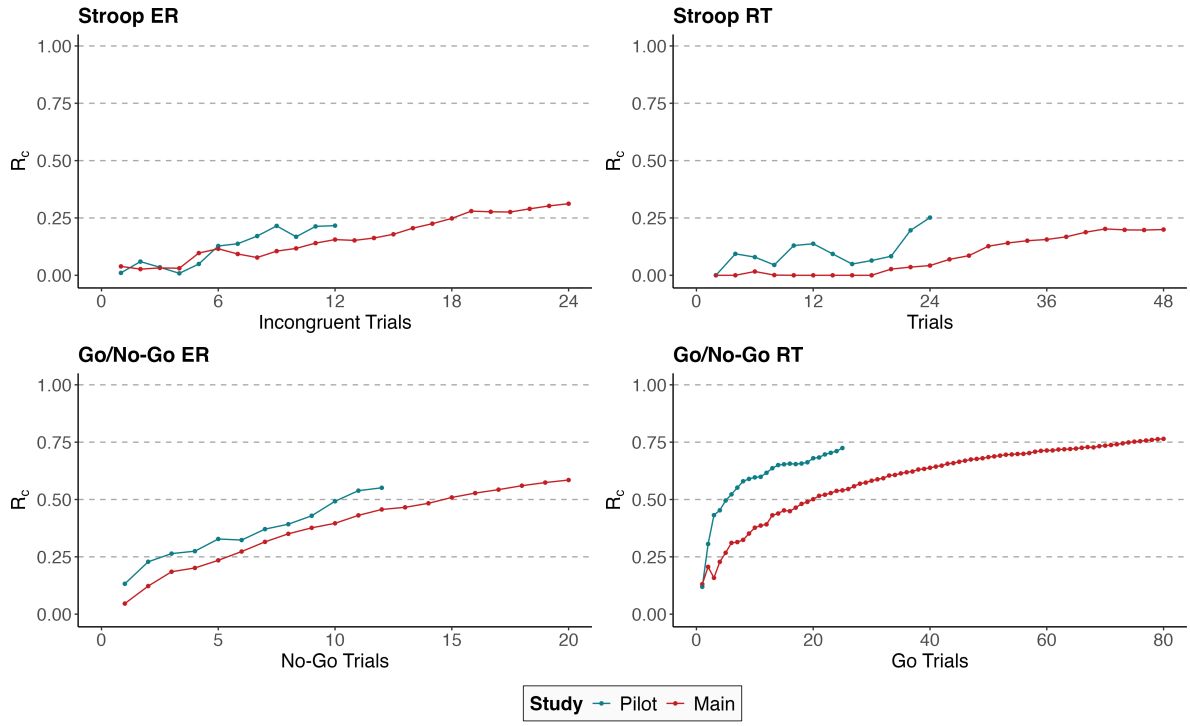
Trend of task scores across the data collection period



Note. ER error rate; RT response time; RT-Diff difference in response times between congruent and incongruent trials.

Figure 5

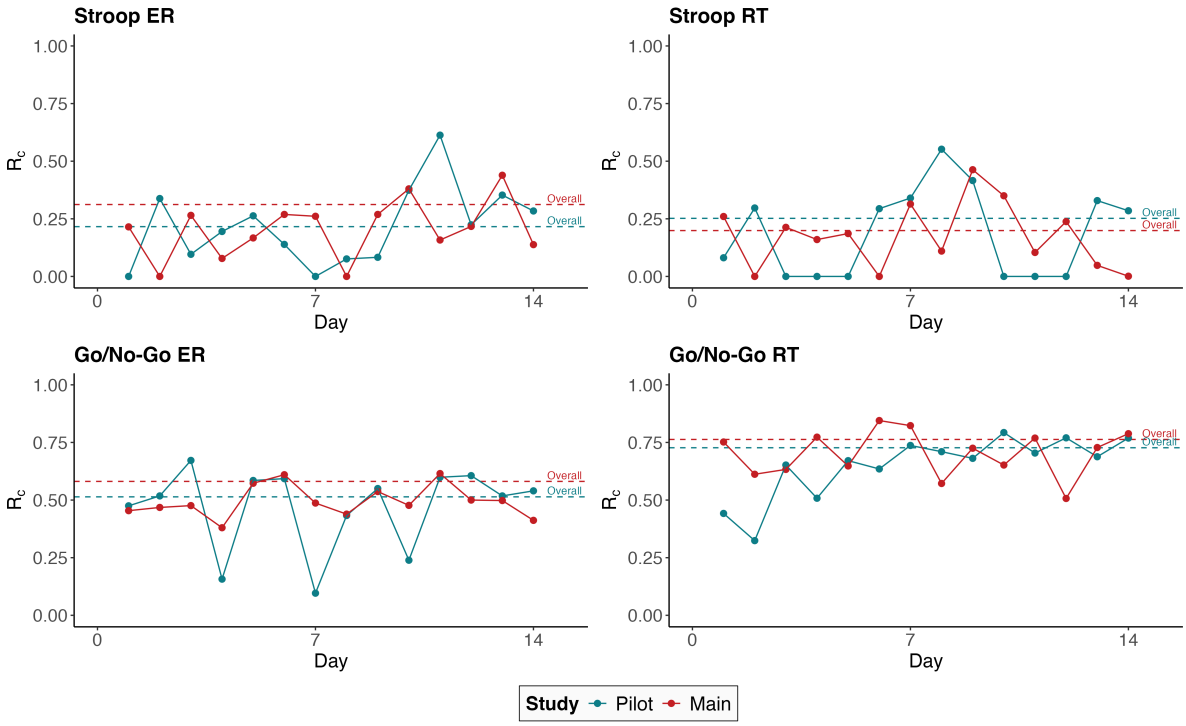
Within-subject reliability with increasing number of trials



Note. ER error rate; RT response time; RT-Diff difference in response times between congruent and incongruent trials; note that the x-axis displays the number of trials per block which is half of the total trials. One block of the Stroop task consisted of 12 congruent and 12 incongruent trials in the Pilot Study and 24 congruent and 24 incongruent trials in the Main Study. One block of the Go/No-Go task consisted of 12 no-go and 28 go trials in the Pilot Study and 20 no-go and 80 go trials in the Main Study.

Figure 6

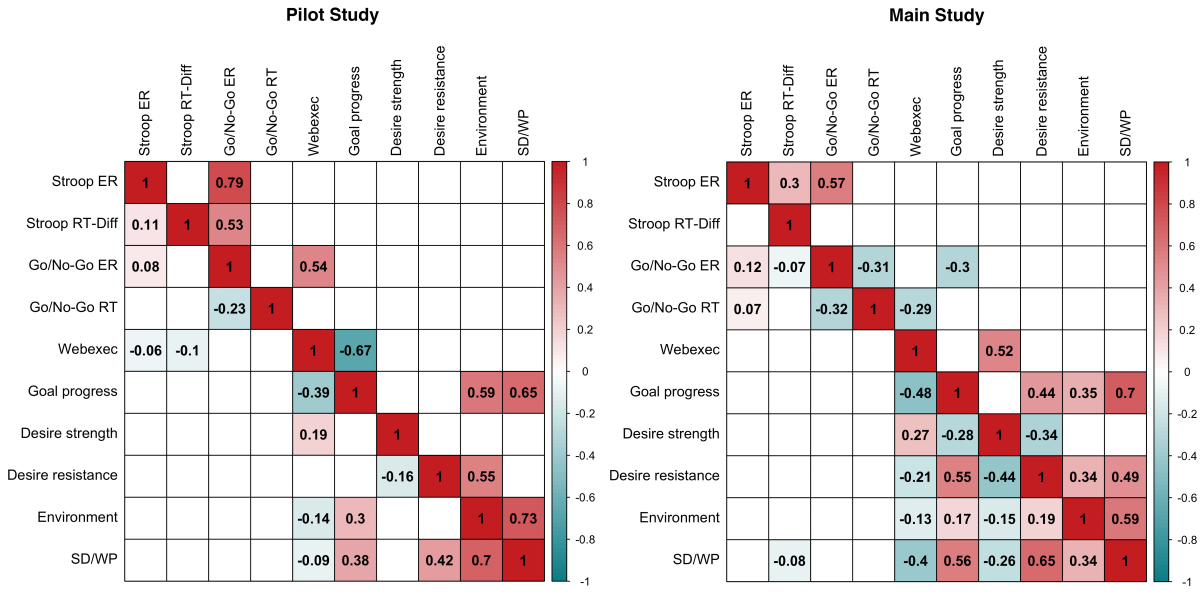
Within-subject reliability for each day of ambulatory assessment



Note. ER error rate; RT response time; RT-Diff difference in response times between congruent and incongruent trials.

Figure 7

Correlation matrix for convergent validity



Note. ER error rate; RT response time; RT-Diff difference in response times between congruent and incongruent trials. Upper-right triangles display between-subject correlations and lower-left triangles display within-subject correlations. Only significant correlations are shown.