Behavior Research Methods (2018) 50:1345–1358
https://doi.org/10.3758/s13428-018-1036-5

Comparability, stability, and reliability of internet-based mental chronometry in domestic and laboratory settings

R. Miller[1,2] · K. Schmidt[1] · C. Kirschbaum[1] · S. Enge[3]

Published online: 15 March 2018

Abstract

The internet-based assessment of response time (RT) and error rate (ERR) has recently become a well-validated alternative to traditional laboratory-based assessment, because methodological research has provided evidence for negligible setting- and setup-related differences in RT and ERR measures of central tendency. However, corresponding data on potential differences in the variability of such performance measures are still lacking, to date. Hence, the aim of this study was to conduct internet-based mental chronometry in both poorly standardized domestic and highly standardized laboratory environments and to compare the variabilities of the corresponding performance measures. Using the Millisecond Inquisit4Web software, 127 men and women completed three different RT-based cognitive paradigms (i.e., go/no-go, two-back, and number–letter). Each participant completed all paradigms in two environments (i.e., at home and in the laboratory), with a time lag of seven days and in a counterbalanced order. Mixed-effects modeling was employed to estimate the between-setting variability across a comprehensive set of performance measures, including conventional measures of central tendency (i.e., mean RT and ERR) and further measures characterizing the joint distribution of RT/ERR. The latter measures were estimated using the diffusion model. The results suggested negligible differences between the domestic and laboratory settings. Thus, this study provides novel evidence suggesting that the statistical power of internet-based mental chronometry is commonly not compromised by increased environmental variance. The within- and between-session reliabilities were in a satisfactory range—that is, comparable to performance measures collected offline in laboratory settings. In consequence, our results support the broad applicability, robustness, and cost efficiency of mental chronometry assessment using the internet.

Keywords Internet · Reliability · Task switching · Inhibition · Updating · Cognitive control · Executive functions · Diffusion model

Theoretical background

The impact of the internet on modern daily life is immense. Thus, it is not surprising that it has also found its way into experimental psychology, with 11%–31% of the studies published in major cognitive science journals relying on internet-based data collection (Stewart, Chandler, & Paolacci, 2017). Over the last few years the systematic investigation of internet-based response time (RT) assessments has provided evidence against most of the preconceptions that initially hindered their broad application (Germine et al., 2012). Therefore, we know that mental chronometry using the internet is almost as precise as the traditional assessments conducted offline in laboratory settings.

To date, a considerable number of studies have investigated absolute RTs collected online. Keller, Gunasekharan, Mayo, and Corley (2009) evaluated the precision of internet-based recordings of known time intervals and found it to be remarkably high (up to a 22-ms offset in Windows operating systems). However, data based on human RTs usually indicate a small degree of overestimation in internet-assessed RT measures (i.e., 10–100 ms; Brand & Bradley, 2012) as compared

R. Miller and K. Schmidt contributed equally to this work.

R. Miller
robert.miller@tu–dresden.de

K. Schmidt
kornelius.schmidt@tu–dresden.de

[1] Faculty of Psychology, Technische Universität Dresden, Dresden, Germany

[2] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

[3] Faculty of Natural Sciences, Medical School Berlin, Berlin, Germany

to offline measures. Nonetheless, this systematic overestimation of internet-assessed RT data is usually negligible (e.g., Chetverikov & Upravitelev, 2016; Germine et al., 2012; Reimers & Stewart, 2007) and, therefore, hardly compromises the replicability of most cognitive paradigms. Crump, McDonnell, and Gureckis (2013) conducted an internet-based replication of eight widely used cognitive paradigms (e.g., Stroop, switching, flanker, Simon, and attentional blink). In this study, only the masked-priming effect could not be successfully replicated, due to a lack of precision with presentation times shorter than 65 ms. An extensive body of evidence exists to support the successful online replication of cognitive paradigms, such as the studies of Simcox and Fiez (2014; replication of a flanker effect and the lexical-decision paradigm), Barnhoorn, Haasnoot, Bocanegra, and van Steenbergen (2015; replication of the Stroop effect, attentional blink effect, and masked-priming effect), and Reimers and Maylor (2005; replication of the task-switching paradigm).

In light of these findings, internet-based mental chronometry appears to have a significant advantage: Cognitive paradigms can be presented to a wide range of participants while collecting data in an instant, with little cost (Reips, 2002). However, performance assessments via the internet seem to come with the drawback of some additional, unexplained variability (e.g., Neath, Earle, Hallett, & Surprenant, 2011; Schubert, Murteira, Collins, & Lopes, 2013). Brand and Bradley (2012) distinguished two putative main areas as sources of this variability—that is, technical variability and environmental variability. It is a well-known fact that the use of different testing hardware (e.g., CPU, keyboard, screen, or mouse) drives some variability in RT data (Neath et al., 2011; Plant & Turner, 2009). The same applies to diverse types of software, such as varying operating systems, driver versions (Plant & Quinlan, 2013), and web browsers (Reimers & Stewart, 2015; Semmelmann & Weigelt, 2017). Added to these technical aspects, the general lack of experimental control (to avoid effect confounding by person and environmental factors) and standardization (to minimize performance variance due to different assessment environments) probably influences internet-based mental chronometry. Thus, the difficulty of controlling for distraction (Brand & Bradley, 2012), as well as the missing guidance and control provided by an experimenter (Reips, 2002), can decrease data quality.

Research rationale

Although many studies in past years have aimed to prove the general precision of internet-based mental chronometry, certain information about the quality of such assessments is still missing. According to Germine et al. (2012), the quality of performance measures is reflected by three main aspects—that is, (1) their central tendency, (2) their variance, and (3) their reliability. These three quality indicators will help us to identify several properties of internet-based mental chronometry that have not been investigated so far.

With regard to the first aspect, most of the above-mentioned studies (e.g., Barnhoorn et al., 2015; Crump et al., 2013; Simcox & Fiez, 2014) focused on differences in the central tendencies of laboratory- versus internet-assessed performance measures. Given that such differences in the central tendencies were largely negligible, the respective paradigms were considered replicable in both settings. Notably, such replicability statements hinge on the type of performance measure. Commonly, the mean or median RT of each individual is considered the primary performance measure for mental chronometry. Recently, however, the joint analysis of RTs and errors (ERR) using the diffusion model (DM; Voss, Nagler, & Lerche, 2013) has gained popularity, because its measures provide additional information based on higher-order moments (e.g., the skewness) of the individual RT distributions (see also Wagenmakers, 2009). Because the central tendencies of such measures have not been compared between laboratory- and internet-based assessments, the present study went beyond the investigation of conventional performance measures by also fitting the DM and replicating its effects in different cognitive paradigms.

With regard to the second aspect, increased variability in internet-assessed performance measures has often been presumed (Neath et al., 2011; Reimers & Stewart, 2015). Nonetheless, we still lack estimates of the practical extent of the variance increase, which has important implications for the statistical power to successfully replicate cognitive paradigms. One attempt to provide such data was carried out by Brand and Bradley (2012), but their estimates were informed only by simulated data and considered the portion of technical (i.e., setup) variability in performance measures. By contrast, the portion of environmental variability was disregarded, although there has been a debate on the latter aspect for almost as long as RTs have been assessed online (Hecht, Oesker, Kaiser, Civelek, & Stecker, 1999). Another attempt to estimate the additional variance introduced by online data collection was carried out by de Leeuw and Motz (2016). However, those authors focused exclusively on software differences in offline versus online assessments, but disregarded actual setting variability. In consequence, in the present study we aimed to further investigate the impact of deviations from standardized laboratory settings on variability in internet-assessed performance measures (including those obtained by diffusion modeling). Detailed insights on this variability will help us quantify the putative loss of statistical power whenever cognitive performance is assessed in domestic environments.

In terms of the number of published studies, the reliabilities of performance measures (i.e., the third aspect) are seldom explicitly quantified in the field of cognitive psychology, although they convey important information about the maximum effect that can be attained in a given paradigm (e.g., Paap & Sawi, 2016). To our knowledge, only Germine et al. (2012) has provided reliability information about some internet-assessed performance measures. However, this information was restricted to the internal consistency of ERR data. Hence, neither the consistency nor the test–retest stability of internet-assessed RT and ERR measures has yet been reported. In the present study we therefore estimated the within- and between-session reliabilities of internet-based mental chronometry in different cognitive paradigms.

Method

Sample

A total of 137 students at Technische Universität Dresden took part in the study. However, only 127 of the participants (33 male, 94 female; between 18 and 40 years, $M_{\mathrm{age}} = 23.6$ years, $SD = 4.1$) completed both sessions and, hence, provided a complete set of data. Informed consent was given by each participant prior to the procedure. Each participant received €18 in compensation or, in the case of psychology students, the equivalent in credit points. Approval was granted by the local ethics committee.

Apparatus and materials

Internet-based paradigm presentation was implemented in both the lab and home settings by Inquisit4Web (Millisecond Software, Seattle, USA), a software for Windows and OSX systems that operates client-sided.

In the lab setting (*lab*), participants used a Windows computer (Windows XP Professional, Version 202, Service Pack 3) in combination with a 19-in. TFT display, a QWERTZ keyboard, and an optical mouse. The online experiment was initiated using the Firefox browser (version 39.0). For the domestic setting (*home*), the only technical restriction given was to use a desktop or laptop PC in combination with a physical mouse and keyboard. No tablets or smartphones were permitted. Responses via a touch-sensitive surface were not allowed.

To allow application of our findings to a broad field of RT research, we aimed to investigate a diverse range of cognitive constructs. According to Miyake et al. (2000), the three interrelated constructs "shifting," "updating," and "inhibition" cover the majority of performance variance in the RT tasks commonly used to assess executive functioning. We chose three RT tasks to tap into each of these constructs: a number–letter task (Rogers & Monsell, 1995; Fig. 1A) to assess "shifting," a go/no-go task (Wolff et al., 2016; Fig. 1B) to assess "inhibition," and a spatial two-back task (Friedman et al., 2008; Fig. 1C) to assess "updating." In each trial of the number–letter task, a character–digit pair was presented below or above a black bar. The participants were asked to classify either the character (*a* or *b*) or the number (*1* or *2*) when the stimulus was presented below or above the bar, respectively. The "y" key (in response to *a* and *1*) and the "m" key (in response to *2* and *b*) on the keyboard served as response keys. Switch and repeat trials were defined by the compatibility of the target (i.e., number or letter) to that in the previous trial. The task consisted of 256 trials total (50% switch trials, 50% repeat trials). The go/no-go task simply required the classification of the alignment of two circles as vertical (go, predominant trial type) or horizontal (no go). Because DM analyses require two-choice data, the classical go/no-go task was slightly modified: Both go trials ("y" key) and no-go trials ("m" key) required a key response. The task consisted of 400 trials—that is, 87.5% of these were go trials, and 12.5% were no-go trials. Hence, the predominant response tendencies established by the go stimuli needed to be inhibited in the rarer no-go trials. The two-back task required continuously determining the matching of the currently presented stimulus array and the stimulus array that had been presented two trials before (target or nontarget), using the same keys that were used in the other paradigms. Hence, the task required successfully updating working memory in order to be correctly performed. The task consisted of 160 trials (70% nontargets, 30% targets).

Procedure

Participants were randomly assigned to the first setting—that is, "lab" or "home". Prior to the first session, each participant had received an email that contained all information about the testing (i.e., the consent form, instructions for using the individualized participation code, and assessment dates), as well as a web link that led to the fully automatized online experiment. After starting the experiment, all participants were asked to enter their participation code and whether the current session was being performed in the lab or at home. Next, they were asked to report their sex and age. Thereafter, the tasks were presented in the following order: number–letter task, go/no-go task, and two-back task. Each task was preceded by task instructions and practice trials. Between tasks, the possibility for a short break was given. The whole procedure took approximately 1 h per session. The time between sessions was instructed to be seven days.

Only in the lab was an experimenter present, who welcomed the participants, guided them to the computer, and dismissed them after the procedure had been completed. Notably, the experimenter did not interact with the participants

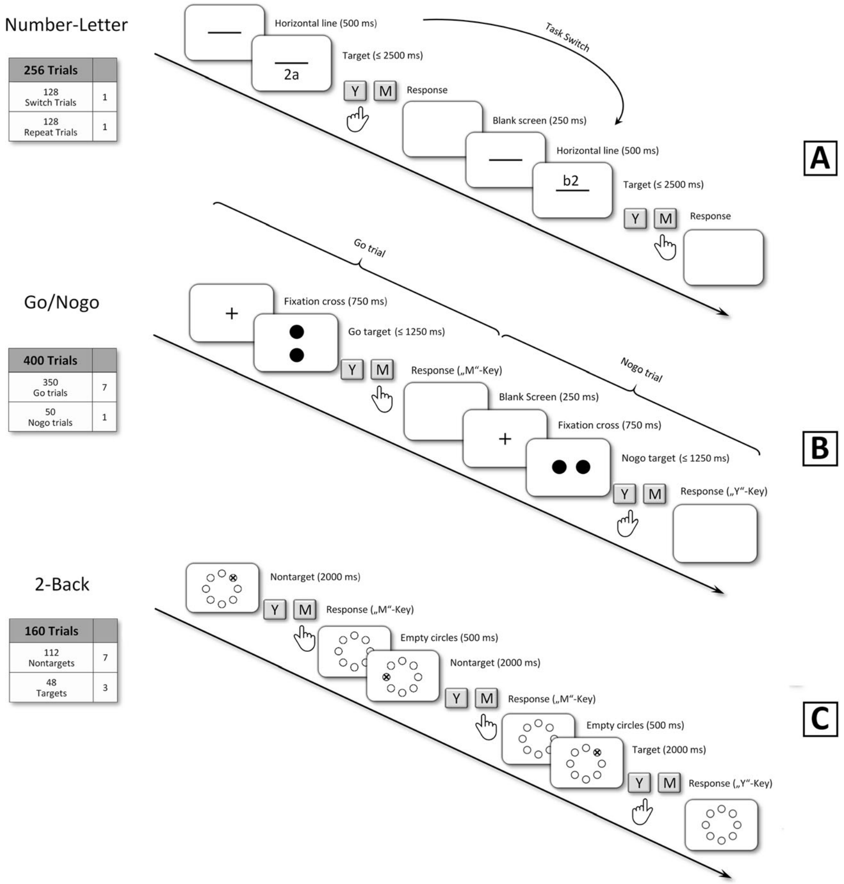

Fig. 1 Illustration of the task battery employed, which consisted of three executive-functioning tasks: (A) number–letter, (B) go/no-go, and (C) spatial two-back

during the administration of the online task battery. For the home session, participants were told to conduct the experiment in a calm and nondistracting environment.

Performance measures

RT and ERR data enable investigation of a broad range of different performance measures. However, we focused on those measures that are commonly used as performance indicators in studies using each of the respective tasks. In the number–letter task, we focused on the difference in mean log(RT) and the difference in the relative error frequencies (ERR) between switch and repeat trials. By contrast, we focused on log(RT) and ERR in the no-go trials of the go/no-go task, as well as in the target trials of the spatial two-back task. Besides these six conventional performance measures, which were of primary interest to our analyses, further, secondary measures are listed in the Appendix.

Lately, the DM has become increasingly popular for joint analyses of RT and ERR data from cognitive two-choice tasks. To account for this development, we also performed DM analyses. The measures of the DM allow for a finer, theory-driven interpretation of the underlying cognitive processes, as compared to conventional performance measures based on aggregate data, such as mean RTs (Voss et al., 2013; Voss, Rothermund, & Voss, 2004). The predominantly estimated measures are upper threshold boundary separation (a), relative starting point (zr), drift rate (v), and response time constant ($t0$). The most prominent measure is v (typical range: $-5 < v < 5$), which describes the mean speed of the information accumulation process toward the correct response option. The higher the boundary separation a is ($0.5 < a < 2$), the more cautious is the response style of the individual. A preference toward one of the two response options is expressed by zr ($.3 < zr < .7$), whereas $zr = .5$ indicates no preference. All residual processes (i.e., sensory encoding and motor execution of the response) are expressed by $t0$ ($.1 < t0 < .5$). For all trials of the respective task, we estimated the common a, zr, and $t0$ measures, whereas v was estimated separately for each trial type of the go/no-go task and the two-back task. In the case of the number–letter task, all DM measures were estimated separately for switch and repeat trials, before performance measures were calculated from their differences between the two conditions. An overview of all analyzed performance measures and their labels is presented in Table 1.

Statistical analyses

All two-stage analyses reported in this article were performed using the R 3.3.1 statistical software (R Core Team, 2017). The raw data and the R script of our analyses can be downloaded from https://osf.io/64q2z/.

To obtain performance measures at the first analysis stage, thorough outlier removal is a common practice in internet-based mental chronometry, to control for unwanted environmental variability, such as transient distractions (Brand & Bradley, 2012; Keller et al., 2009). Because we were specifically interested in these effects, only little outlier control was applied to our data. Prior to the analyses, all timed-out responses were excluded (0.5% of all trials). Furthermore, trials with log(RT) lower or higher than 2.5 standard deviations (*SD*) from the conditional mean RT of each session, task, and trial type were excluded (1.9% of all trials). Moreover, three participants were excluded from the analyses of the two-back task due to technical problems during this paradigm in the home session.

Table 1 Overview of the analyzed performance measures

Label	Paradigm	Measure (*y*)
Conventional measures		
C-NL$_{\text{RT}}$	Number–Letter	$M[\log(\text{RT}_{\text{Switch}})] - M[\log(\text{RT}_{\text{Repeat}})]$
C-NL$_{\text{ERR}}$	Number–Letter	$\text{ERR}_{\text{Switch}} - \text{ERR}_{\text{Repeat}}$
C-GN$_{\text{RT}}$	Go/No-Go	$M[\log(\text{RT}_{\text{No-go}})]$
C-GN$_{\text{ERR}}$	Go/No-Go	ERR_{Nogo}
C-2B$_{\text{RT}}$	Two-Back	$M[\log(\text{RT}_{\text{Target}})]$
C-2B$_{\text{ERR}}$	Two-Back	$\text{ERR}_{\text{Target}}$
Measures from diffusion modeling		
DM-NL$_{\text{a}}$	Number–Letter	$a_{\text{Switch}} - a_{\text{Repeat}}$
DM-NL$_{\text{zr}}$	Number–Letter	$zr_{\text{Switch}} - zr_{\text{Repeat}}$
DM-NL$_{\text{t0}}$	Number–Letter	$t0_{\text{Switch}} - t0_{\text{Repeat}}$
DM-NL$_{\text{v}}$	Number–Letter	$v_{\text{Switch}} - v_{\text{Repeat}}$
DM-GN$_{\text{a}}$	Go/No-Go	$a_{\text{Go}} + {}_{\text{No-go}}$
DM-GN$_{\text{zr}}$	Go/No-Go	$zr_{\text{Go}} + {}_{\text{No-go}}$
DM-GN$_{\text{t0}}$	Go/No-Go	$t0_{\text{Go}} + {}_{\text{No-go}}$
DM-GN$_{\text{v}}$	Go/No-Go	$v_{\text{Go}} - v_{\text{No-go}}$
DM-2B$_{\text{a}}$	Two-Back	$a_{\text{Target}} + {}_{\text{Nontarget}}$
DM-2B$_{\text{zr}}$	Two-Back	$zr_{\text{Target}} + {}_{\text{Nontarget}}$
DM-2B$_{\text{t0}}$	Two-Back	$t0_{\text{Target}} + {}_{\text{Nontarget}}$
DM-2B$_{\text{v}}$	Two-Back	$v_{\text{Target}} - v_{\text{Nontarget}}$

Subscripts in the measure (*y*) column denote the respective trial type of the cognitive paradigm. RT = response time; ERR = relative error frequency; *a* = boundary separation (response caution); *zr* = relative starting point (response bias); *t*0 = response time constant (nondecision time); *v* = drift rate (speed of evidence accumulation)

For each task, the above-mentioned performance measures were calculated and then submitted to the second analysis stage. DM measures were estimated by minimizing the Kolmogorov–Smirnov statistic between the observed and the model-implied distribution of correct and error RTs using fast-dm (Voss & Voss, 2007).

At the second analysis stage, hierarchical regression analyses of each measure were conducted using generalized least squares estimation. The conditional mean of the respective performance measure (*y*) was modeled by the intercept (β_0), the setting (β_1; home = 0, lab = 1), the number of the session (β_2; 1st session = 0, 2nd session = 1), and the setting of the initial session (β_3; home = 0, lab = 1). Accordingly, the structural part of the regression model can be expressed by Formula 1:

$$y = \beta_0 + \beta_1 + \beta_2 + \beta_{3.} \tag{1}$$

Formal hypothesis testing for setting differences (β_1) in the 18 primary performance measures was performed on the basis of a tail-area false discovery rate of FDR = 5% (Benjamini & Hochberg, 1995).

The employed regression models also accounted for the difference between the residual standard deviation (*SD*) of *y* in the lab (σ) as compared to home ($\omega * \sigma$). The resulting difference in the variability of performance measures between lab and home was therefore expressed as an *SD* ratio (ω), whereas the *SD* in the lab (σ) was set as the reference. Hence, the following formula (Eq. 2) applied:

$$\omega = SD_{\text{Home}}/SD_{\text{Lab}} = SD_{\text{Home}}/\sigma. \tag{2}$$

Each regression model further provided estimates of the correlation between the two sessions (r_{TR}). Given the design of the present study, the correlation coefficient *r* was interpretable as an estimate of the test–retest stability (time lag: 1 week) of the respective performance measures. To provide references for these stability estimates, the internal reliability of each performance measure was also quantified by splitting between odd- and even-numbered trials and correcting these estimates for attenuation by using the Spearman–Brown formula.

To obtain confidence intervals (CIs) for the estimated *SD* ratios (ω), test–retest stabilities (r_{TR}), and internal reliabilities (r_{Lab} and r_{Home}), a nonparametric bootstrap with n = 50,000 replicates was performed for each model.

Results

To provide a general description of the investigated performance measures, Table 2 lists their means and standard

Table 2 Mean performance measures and their mean standard deviations across participants

Performance Measure	Lab		Home	
	M	*SD*	*M*	*SD*
Conventional measures				
C-NL$_{RT}$	0.268	0.099	0.261	0.111
C-NL$_{ERR}$	0.031	0.039	0.027	0.039
C-GN$_{RT}$	6.000	0.142	6.000	0.162
C-GN$_{ERR}$	0.192	0.134	0.208	0.159
C-2B$_{RT}$	6.573	0.244	6.526	0.220
C-2B$_{ERR}$	0.134	0.117	0.126	0.130
Measures from diffusion model				
DM-NL$_a$	– 0.192	0.417	– 0.154	0.384
DM-NL$_{zr}$	– 0.128	0.145	– 0.135	0.151
DM-NL$_{t0}$	0.194	0.114	0.170	0.103
DM-NL$_v$	0.010	0.521	0.026	0.543
DM-GN$_a$	1.457	0.426	1.414	0.370
DM-GN$_{zr}$	0.526	0.095	0.523	0.113
DM-GN$_{t0}$	0.204	0.037	0.207	0.045
DM-GN$_v$	2.499	1.887	2.462	2.104
DM-2B$_a$	1.737	0.356	1.711	0.338
DM-2B$_{zr}$	0.363	0.089	0.381	0.097
DM-2B$_{t0}$	0.400	0.107	0.384	0.094
DM-2B$_v$	1.161	0.980	0.980	1.080

M = mean across participants; *SD* = mean standard deviation across participants. The labeling of performance measures according to Table 1

deviations in both settings (i.e., lab vs. home). The inferential analyses are based on multiple hierarchical regression modeling to compare the performance measures between the different settings. A comprehensive list of all estimates is provided in Table 3. On the basis of the Benjamini–Hochberg procedure (FDR = 5%), $p \leq .02$ was considered the significance threshold for formal hypothesis testing.

Systematic setting differences

Regression modeling yielded no evidence of significant differences between the settings (i.e., lab and home), for both the conventional performance measures ($-0.01 \leq \beta_1 \leq 0.01$, $.24 \leq p \leq .61$) and the DM-based measures ($-0.04 \leq \beta_1 \leq 0.20$, $.05 \leq p \leq .81$).

The change in performance measures from the first to the second session was estimated by β_2 (see Table 3). Some models—that is, C-GN$_{RT}$, C-GN$_{ERR}$, C-2B$_{RT}$, C-2B$_{ERR}$, DM-NL$_{t0}$, DM-GN$_{t0}$, DM-GN$_v$, DM-2B$_a$, and DM-2B$_{t0}$—suggested improvements in task performance ($p \leq .01$). This probably reflects the well-known practice effects that occur in repeatedly conducted cognitive paradigms (e.g., Davidson, Zacks, & Williams, 2003; Enge et al., 2014). All remaining estimates of β_2 provided no considerable evidence of performance improvements ($.08 \leq p \leq .87$).

To control for potential asymmetries in the initial condition, β_3 estimated differences in performance measures due to the initial setting. Models DM-NL$_{zr}$, DM-NL$_{t0}$, DM-NL$_v$, and DM-2B$_{t0}$ revealed borderline associations ($.01 \leq p \leq .04$), whereas all other models suggested that differences due to initial setting were negligible ($.21 \leq p \leq .99$).

Variability in the two settings

Each regression model quantified the differences in residual variability of the 18 performance measures between the two settings. The estimated *SD* ratio ω quantified the relative change of *SD* in the home condition as compared to the lab condition (σ). The majority of models revealed a $\omega > 1$ ($1.01 \leq \omega \leq 1.24$), indicating higher variability of the performance measures at home (see Fig. 2). However, seven out of the 18 models—that is, C-NL$_{ERR}$, C-2B$_{RT}$, DM-NL$_a$, DM-NL$_{t0}$, DM-GN$_a$, DM-2B$_a$, and DM-2B$_{t0}$—showed a $\omega < 1$ ($0.86 \leq \omega \leq 0.98$). In some cases the variability of performance measures might therefore have been smaller in the domestic setting than in the standardized lab setting.

Bootstrapping was used to estimate the sampling variability (i.e., the 95% confidence intervals [CIs]) of all the estimated ω values. According to these analyses, only two models showed an ω that differed considerably from 1. Model C-GN$_{ERR}$, representing the ERR of no-go trials in the no/no-go task, yielded $\omega = 1.19$ [1.02, 1.39]. Model DM-GN$_{zr}$, estimating *zr* in the go/no-go task, yielded $\omega = 1.20$ [1.05, 1.36]. Both models provided evidence that the variability of some performance measures may have increased in the less standardized setting.

To increase the precision of the estimated ωs, we finally pooled the performance measures of each task through Bayesian meta-analyses (while accounting for the between-measure variability of ω using the Berger–Bernardo reference prior; see Bodnar, Link, Arendacká, Possolo, & Elster, 2017). In the go/no-go task, the mean *SD* ratio increased by 14.9% in the domestic setting ($\omega = 1.15$, CI$_{95\%}$ = 0.96–1.37). By contrast, the number–letter and two-back tasks yielded no considerable evidence for such an increase, with $\omega = 1.00$ (CI$_{95\%}$ = 0.83–1.20) and $\omega = 1.01$ (CI$_{95\%}$ = 0.84–1.23), respectively. Note that ω did not differ systematically between the conventional and DM-based performance measures. A precise record of all estimated ωs and their CIs is given in Table 3. Additional calculations using further performance measures confirmed these findings and are provided in the Appendix.

Reliability in both settings

The test–retest stability r_{TR} (time lag: seven days) between the two sessions was also estimated by the regression models. For

Table 3 Regression estimates with variance estimation according to settings and the reliabilities of the respective performance measures

Outcome	$\beta_0 \pm SE$ (p)	$\beta_1 \pm SE$ (p)	$\beta_2 \pm SE$ (p)	$\beta_3 \pm SE$ (p)	σ	ω [95% CI$_\omega$]	r_{TR} [95% CI$_r$]	r_{Lab} [95% CI$_r$]	r_{Home} [95% CI$_r$]
C-NL$_{RT}$	0.25 ± 0.02 (<.001)	0.01 ± 0.01 (.31)	0.00 ± 0.01 (.87)	0.02 ± 0.02 (.37)	0.10	1.12 [0.97, 1.29]	.69 [.58, .77]	.85 [.79, .89]	.91 [.88, .93]
C-NL$_{ERR}$	0.02 ± 0.01 (<.001)	0.00 ± 0.00 (.24)	0.00 ± 0.00 (.27)	0.01 ± 0.01 (.24)	0.04	0.98 [0.83, 1.19]	.33 [.13, .52]	.33 [– .02, .55]	.35 [.09, .54]
C-GN$_{RT}$	6.04 ± 0.02 (<.001)	– 0.01 ± 0.01 (.26)	– 0.05 ± 0.01 (<.001)	– 0.02 ± 0.03 (.41)	0.14	1.12 [0.98, 1.28]	.73 [.64, .79]	.92 [.88, .95]	.92 [.88, .95]
C-GN$_{ERR}$	0.19 ± 0.02 (<.001)	– 0.01 ± 0.01 (.42)	0.03 ± 0.01 (<.01)	– 0.01 ± 0.02 (.81)	0.13	1.19 [1.02, 1.39]	.61 [.49, .71]	.85 [.78, .90]	.80 [.70, .87]
C-2B$_{RT}$	6.62 ± 0.03 (<.001)	0.01 ± 0.02 (.54)	– 0.19 ± 0.02 (<.001)	0.03 ± 0.04 (.32)	0.22	0.95 [0.84, 1.09]	.67 [.54, .77]	.96 [.94, .97]	.94 [.92, .96]
C-2B$_{ERR}$	0.14 ± 0.02 (<.001)	0.00 ± 0.01 (.61)	– 0.03 ± 0.01 (<.001)	0.00 ± 0.02 (.90)	0.12	1.12 [0.87, 1.44]	.63 [.48, .76]	.84 [.73, .90]	.87 [.78, .92]
DM-NL$_a$	– 0.17 ± 0.05 (<.01)	– 0.03 ± 0.05 (.51)	0.02 ± 0.05 (.65)	0.01 ± 0.05 (.89)	0.42	0.92 [0.78, 1.08]	.04 [– .19, .26]	.01 [– .39, .31]	.38 [.16, .55]
DM-NL$_{zr}$	– 0.18 ± 0.02 (<.001)	0.01 ± 0.02 (.53)	0.02 ± 0.02 (.20)	0.05 ± 0.02 (<.01)	0.14	1.01 [0.85, 1.20]	.25 [.09, .39]	.36 [.05, .57]	.47 [.24, .63]
DM-NL$_{t0}$	0.17 ± 0.01 (<.001)	0.02 ± 0.01 (.12)	– 0.03 ± 0.01 (.01)	0.03 ± 0.02 (.04)	0.11	0.93 [0.76, 1.14]	.30 [.13, .45]	.48 [.25, .65]	.61 [.43, .74]
DM-NL$_v$	0.20 ± 0.07 (<.01)	– 0.04 ± 0.06 (.54)	– 0.11 ± 0.06 (.08)	– 0.19 ± 0.07 (.01)	0.52	1.01 [0.85, 1.20]	.18 [.03, .34]	.34 [.05, .54]	.47 [.25, .62]
DM-GN$_a$	1.47 ± 0.05 (<.001)	0.03 ± 0.04 (.51)	– 0.07 ± 0.04 (.12)	– 0.02 ± 0.06 (.67)	0.43	0.86 [0.62, 1.21]	.23 [.08, .43]	.57 [.38, .73]	.63 [.47, .82]
DM-GN$_{zr}$	0.52 ± 0.02 (<.001)	0.00 ± 0.01 (.81)	0.00 ± 0.01 (.62)	0.00 ± 0.02 (.88)	0.09	1.20 [1.05, 1.36]	.54 [.40, .67]	.77 [.68, .84]	.77 [.66, .85]
DM-GN$_{t0}$	0.21 ± 0.01 (<.001)	– 0.01 ± 0.00 (.15)	– 0.01 ± 0.00 (<.01)	0.00 ± 0.01 (.91)	0.04	1.24 [0.98, 1.51]	.51 [.37, .63]	.91 [.87, .95]	.95 [.91, .97]
DM-GN$_v$	2.19 ± 0.29 (<.001)	0.14 ± 0.18 (.43)	0.54 ± 0.18 (<.01)	– 0.09 ± 0.31 (.76)	1.87	1.12 [0.96, 1.32]	.50 [.33, .64]	.85 [.79, .89]	.82 [.74, .88]
DM-2B$_a$	1.82 ± 0.05 (<.001)	– 0.01 ± 0.03 (.61)	– 0.21 ± 0.03 (<.001)	0.02 ± 0.05 (.73)	0.34	0.96 [0.83, 1.11]	.56 [.43, .67]	.83 [.75, .88]	.79 [.71, .87]
DM-2B$_{zr}$	0.37 ± 0.01 (<.001)	– 0.01 ± 0.01 (.13)	0.01 ± 0.01 (.15)	0.00 ± 0.01 (.88)	0.09	1.10 [0.93, 1.29]	.35 [.19, .51]	.62 [.47, .73]	.68 [.54, .78]
DM-2B$_{t0}$	0.39 ± 0.01 (<.001)	0.01 ± 0.01 (.37)	– 0.04 ± 0.01 (<.001)	0.03 ± 0.01 (.04)	0.10	0.93 [0.73, 1.16]	.39 [.22, .54]	.87 [.81, .91]	.87 [.76, .93]
DM-2B$_v$	0.90 ± 0.15 (<.001)	0.20 ± 0.10 (.05)	0.09 ± 0.10 (.35)	0.03 ± 0.16 (.84)	0.98	1.10 [0.88, 1.37]	.43 [.28, .58]	.53 [.24, .72]	.71 [.57, .81]

CI = confidence interval; p = p value; r_{Home} = internal reliability at home; r_{Lab} = internal reliability in the lab; r_{TR} = test–retest stability (lag: seven days), β_0 = intercept; β_1 = setting (home = 0, lab = 1); β_2 = number of session (1st session = 0, 2nd session = 1); β_3 = initial setting (home = 0, lab = 1); σ = standard deviation (*SD*) in the lab; ω = *SD* ratio with σ as reference; *SE* = standard error; labeling of performance measures according to Table 1

the conventional measures the test–retest stabilities fell within the range $.33 \leq r_{TR} \leq .73$. The DM-based measures generally showed lower test–retest stabilities—that is, $.04 \leq r_{TR} \leq .56$ (see Fig. 3). All test–retest stabilities and their CIs are listed in Table 3.

To provide benchmarks for r_{TR}, the internal reliabilities (i.e., r_{Lab} and r_{Home}) were subsequently estimated for each performance measure (see Table 3). Regarding the conventional measures, internal reliability had the ranges $.33 \leq r_{Lab} \leq .96$ in the lab and $.35 \leq r_{Home} \leq .94$ at home. The internal reliabilities for DM measures fell within the ranges $.01 \leq r_{Lab} \leq .91$ in the lab and $.38 \leq r_{Home} \leq .95$ at home, with the reliabilities for the number–letter task seeming to be generally lower than those for the other two tasks (see Fig. 3).

Replicability of cognitive effects

Multiple one-tailed Welch tests were conducted to assess the presence of the respective effects of the cognitive tasks. For each combination of task and setting, the mean RTs and mean ERRs were compared between trial types (see Fig. 4). All comparisons yielded $ps \leq .001$ for the task-switching effect, the response inhibition effect, and working memory load, and thus exceeded the significance threshold. The effect sizes were

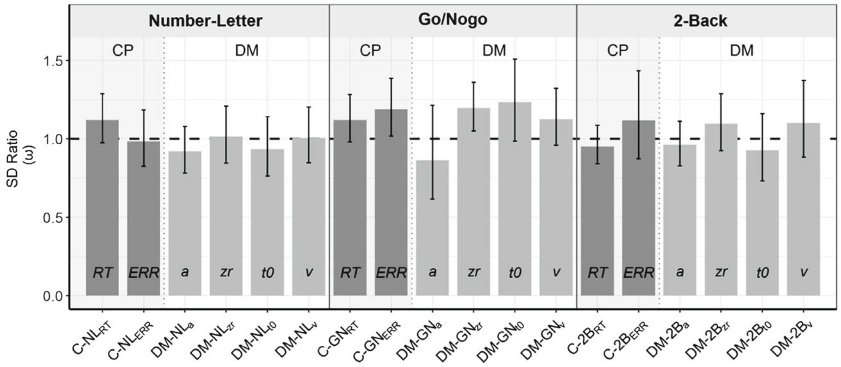

Fig. 2 *SD* ratios ω (home/lab) of the respective performance measures. Error bars indicate the 95% confidence intervals based on bootstrapping (n = 50,000). *SD* ratios $\omega > 1$ indicate more residual variance at home, whereas $\omega < 1$ indicate more residual variance in the lab. CP = conventional measures; DM = measures from the diffusion model; RT = response time; ERR = error rate; a = boundary separation; zr = relative starting point; $t0$ = nondecision time; v = drift rate

scattered in the range $0.70 \le d \le 2.72$ (see Fig. 4). These findings indicate that the prominent cognitive effects from the conventional measures were successfully replicable in both the laboratory and domestic settings.

To replicate previously published DM analyses of task switching, Welch tests were also used to investigate the differences between switch and repeat trials with regard to a, $t0$, and v in both settings. Neither the boundary separations a nor the drift rates v differed significantly between the two trial types in the lab ($ps \ge .07$) or at home ($ps \ge .03$). By contrast, the nondecision time $t0$ increased notably in switch trials across both settings ($p < .001$).

Discussion

The aim of this study was to get a better understanding of data quality when mental chronometry is performed in domestic environments via the internet. To answer this question, we focused on three aspects: setting-related differences in the central tendency of different performance measures (including conventional and DM-based measures), their variability, and their reliability.

Systematic differences in performance measures

With regard to conventional performance measures—that is, aggregated RT and ERR—our results are consistent with those from other studies replicating the cognitive effects of different internet-based cognitive tasks (e.g., Crump et al., 2013). For all three presented paradigms we were able to replicate the known effects in both investigated settings, in the lab and at home.

Regarding the DM measures, it is difficult to compare the reported results with previously published findings. To the best of our knowledge, no DM analyses of the spatial two-back task have been published so far. Conversely, Gomez, Ratcliff, and Perea (2007) have extensively discussed how to fit the go/no-go task using the diffusion model. However, their findings of a bias z toward the go response are (in our experience) not universally accepted, because the go/no-go task is commonly regarded as a one-choice task (see R. Miller, Scherbaum, Heck, Goschke, & Enge, 2017). For the task-switching paradigm, Schmitz and Voss (2012) reported a decrease of the drift rate v and an increase of the nondecision time $t0$ in switch trials. The boundary separation a, however, was insensitive to

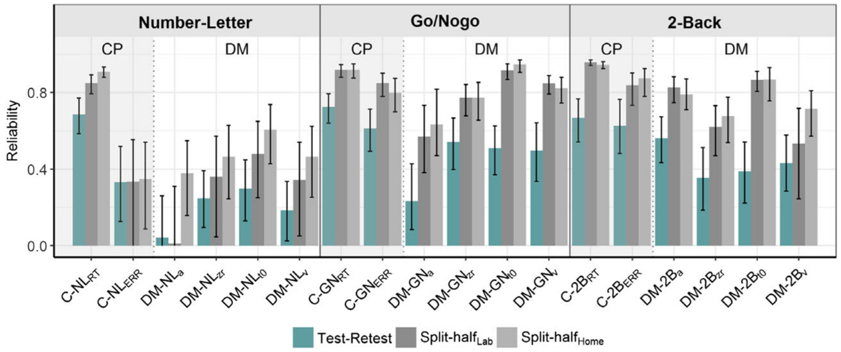

Fig. 3 Test–retest stabilities (interval: seven days) and internal reliabilities of the respective performance measures in both settings. Confidence intervals below 0 are truncated

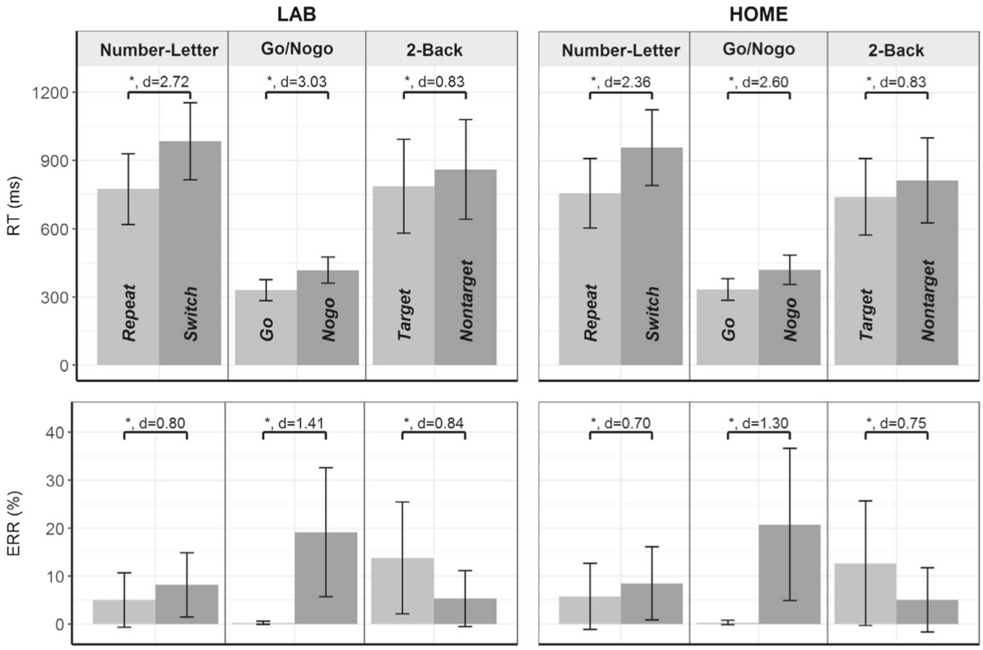

Fig. 4 Mean RTs (upper panels) and mean ERR (lower panels) according to setting (i.e., lab and home) for all tested tasks. Error bars indicate *SD*s. *d* = standardized mean change. $^{*}p < .001$

the trial type. Our analyses replicated these differences for *t*0 (but not after multiplicity adjustment for *v*) in both settings. Similarly, the boundary separation *a* did not differ between the trial types.

Notably, we did not find a meaningful setting difference in the conventional performance measures. This is especially of interest with respect to ERR, because online response logging is supposedly less sensitive to technical variance. Hence, increased ERR can be regarded as an indicator of environmental influences, such as distraction (Semmelmann & Weigelt, 2017). The loss of standardization and experimental control in domestic assessments often goes hand in hand with the preconception that distraction is increased and diligence is decreased (e.g., Chetverikov & Upravitelev, 2016; Gosling, Vazire, Srivastava, & John, 2004; Hilbig, 2016). Yet task completion in the domestic setting had no considerable influence on ERR. These results are consistent with the previous findings of Semmelmann and Weigelt (2017). However, apart from investigating these conventional performance measures, we performed diffusion modeling of the internet-assessed data. Overall, the DM measures showed a pattern similar to that of the conventional measures when lab and home performance was compared. Considering the large size of the investigated per-protocol analysis set ($N = 127$), we can therefore conclude that small effects due to setting differences are unlikely to occur (with a statistical power of 90% for any $d > 0.32$).

In sum, the general comparison of lab and home environments suggested that the relative loss of standardization and environmental control in the domestic setting did not systematically or substantially influence measures of cognitive performance.

Variability of the performance measures

One major aim of this study was to estimate changes in the variability of performance measures that might be caused by domestic data assessment. In general, we observed a slight increase in variability for most of the performance measures at home (5% larger variance, which was primarily attributable to the go/no-go task). However, several aspects need to be pointed out.

Although it seems obvious to expect a smaller variability in the laboratory settings, some of our investigated measures also suggested numerically lower variance in the less standardized and controlled domestic setting. This pattern was found in seven out of 18 reported measures across all three paradigms, and pertained to both the conventional measures (i.e., aggregated RT/ERR data) and the DM measures. The fact that in all of these cases the CIs of the reported *SD* ratios included $\omega = 1$ suggests that the lower *SD*s at home did not reach a practically considerable extent and, therefore, most likely only occurred due to chance. Nevertheless, these results highlight that the variability differences due to a loss of standardization and experimental control might be smaller than is usually expected. Similar patterns of decreased performance variability in the domestic setting were also found in the analyses of additional performance measures (see the Appendix).

Only two out of 18 measures displayed a likely increased setting-related variability. This was the case for ERR and *zr* in the go/no-go task. However, in both cases the lower limit of the 95% CI was very close to $\omega = 1$. Furthermore, both measures only displayed (to our mind) relatively negligible differences ($d \leq .13$).

Brand and Bradley's (2012) simulation study illustrates that the increased variability in unstandardized settings can be compensated for by larger sample sizes. This idea directly results from the logic underlying the estimation of statistical power, which requires the smallest detectable effect of interest to be scaled on the basis of its variability within the investigated sample. Thus, larger sample sizes are needed to achieve the same power level whenever variability increases. For example, a 5% increase in variance would increase the total sample size by 20.5% in order to detect the same performance difference between two participant groups of equal size with the same probability. Thus, informed researchers should carefully evaluate whether their study would actually benefit from the convenience of recruiting large samples online.

Data reliability

First and foremost, no major differences in internal reliability were observed between the two settings. This suggests that the setting was not influential on internal reliability. In general, the internal reliabilities of performance measures for the go/no-go and two-back tasks were in a satisfactory range (i.e., moderate to good). By contrast, the number–letter task seemed to display generally lower (i.e., poor) reliabilities for most of the investigated measures. These findings align well with the internal reliabilities of performance measures that have been derived from similar tasks using laboratory assessment. Wolff et al. (2016), for example, reported internal reliabilities of inverse efficiency scores around $r = .80$ (go/no-go task), $r = .89$ (two-back task), and $r = .59$ (number–letter task). Interestingly, the internal reliabilities also did not seem to differ much between the conventional and DM measures.

Second, and less surprising, the test–retest stabilities (time lag: seven days) were considerably lower than the internal reliabilities. Moreover, we observed a small tendency among the conventional performance measures to reach slightly higher test–retest stabilities than the DM measures (cf. Lerche & Voss, 2017). The large variability of the test–retest stabilities we reported conforms to results from Willoughby and Blair (2011), who compiled similar findings for different executive-functioning tasks. They stated that conventional measures (from laboratory assessments) display reliabilities in the range $.4 \leq r \leq .7$. In the present study, only the ERR in the number–letter task showed a slightly lower reliability. Lerche and Voss found the test–retest stabilities for different DM parameters to be roughly $.0 < r < .8$ (based on optimization of the Kolmogorov–Smirnov statistic), which covers the range of the herein-reported DM measures. Similar to the internal reliabilities, the test–retest stabilities were lower in the number–letter task than in the other two tasks, which can probably be attributed to the way the performance measures from the number–letter task were calculated as compared to those from the other tasks: It has been repeatedly shown and discussed that difference scores are prone to lower reliabilities in many cases (J. Miller & Ulrich, 2013; Paap & Sawi, 2016).

Limitations

When generalizing our findings, two aspects need to be pointed out. First, data collection was achieved by using the Millisecond Inquisit4Web. This Java-based software prioritizes task presentation and thereby minimizes any distraction by background applications. Thus, the reported findings will not necessarily generalize to less invasive presentation software, such as Adobe Flash or JavaScript (e.g., de Leeuw & Motz, 2016; Reimers & Stewart, 2015). Second, our study only assessed undergraduate student participants. Hence, the modulatory impact of a broader education and age range will need to be clarified through further investigations.

Conclusion

The findings of this study provide new insights into the data quality of internet-based mental chronometry in different settings. We were able to show that conventional performance measures as well as DM-based measures are little influenced by unobserved setting variables. Similarly, our findings show that setting-related differences in the variability of performance measures are probably of small size. Although there is an overall increase of variance of approximately 5% in self-chosen experimental environments, internet-based assessments can nevertheless be used without a corresponding loss of statistical power because they facilitate the recruitment of larger sample sizes. Finally, our data indicate that the internal as well as the test–retest stability of internet-assessed cognitive performance is in a satisfactory range and is comparable to reliabilities in the laboratory. This supports the utility and precision of internet-based assessment of cognitive performance in domestic settings.

Author note This work was supported by the German Research Foundation (DFG, Grant No. SFB 940/2).

Appendix

Table 4 Additional performance measures

Label	Paradigm	Measure (y)
Conventional measures		
C-1_A	Number–Letter	$M[\log(\mathrm{RT}_{\mathrm{Repeat}})]$
C-2_A	Number–Letter	$M[\log(\mathrm{RT}_{\mathrm{Switch}})]$
C-3_A	Number–Letter	$\mathrm{ERR}_{\mathrm{Repeat}}$
C-4_A	Number–Letter	$\mathrm{ERR}_{\mathrm{Switch}}$
C-5_A	Go/No-Go	$M[\log(\mathrm{RT}_{\mathrm{Go}})]$
C-6_A	Go/No-Go	$M[\log(\mathrm{RT}_{\mathrm{No\text{-}go}})] - M[\log(\mathrm{RT}_{\mathrm{Go}})]$
C-7_A	Go/No-Go	$\mathrm{ERR}_{\mathrm{Go}}$
C-8_A	Go/No-Go	$\mathrm{ERR}_{\mathrm{No\text{-}go}} - \mathrm{ERR}_{\mathrm{Go}}$
C-9_A	Two-Back	$M[\log(\mathrm{RT}_{\mathrm{Nontarget}})]$
C-10_A	Two-Back	$M[\log(\mathrm{RT}_{\mathrm{Target}})] - M[\log(\mathrm{RT}_{\mathrm{Nontarget}})]$
C-11_A	Two-Back	$\mathrm{ERR}_{\mathrm{Nontarget}}$
C-12_A	Two-Back	$\mathrm{ERR}_{\mathrm{Target}} - \mathrm{ERR}_{\mathrm{Nontarget}}$
Measures from diffusion model		
DM-1_A	Number–Letter	a_{Repeat}
DM-2_A	Number–Letter	zr_{Repeat}
DM-3_A	Number–Letter	$t0_{\mathrm{Repeat}}$
DM-4_A	Number–Letter	v_{Repeat}
DM-5_A	Number–Letter	a_{Switch}
DM-6_A	Number–Letter	zr_{Switch}
DM-7_A	Number–Letter	$t0_{\mathrm{Switch}}$
DM-8_A	Number–Letter	v_{Switch}
DM-9_A	Go/No-Go	v_{Go}
DM-10_A	Go/No-Go	$V_{\mathrm{No\text{-}go}}$
DM-11_A	Two-Back	$v_{\mathrm{Nontarget}}$
DM-12_A	Two-Back	v_{Target}

y = independent variable in regression model; RT = response time; ERR = error rate; a = boundary separation; zr = relative starting point; $t0$ = response time constant; v = drift rate

Table 5 Additional regression estimates with variance estimation according to setting and reliabilities of the respective performance measures

Outcome	$\beta_0 \pm SE$ (p)	$\beta_1 \pm SE$ (p)	$\beta_2 \pm SE$ (p)	$\beta_3 \pm SE$ (p)	σ	ω [95% CI$_\omega$]	r_{TR} [95% CI$_r$]	r_{Lab} [95% CI$_r$]	r_{Home} [95% CI$_r$]
Conventional measures									
C-1_A	6.65 ± 0.02 (<.001)	– 0.01 ± 0.01 (.43)	– 0.15 ± 0.01 (<.001)	– 0.02 ± 0.03 (.44)	0.18	0.96 [0.87, 1.06]	.81 [.74, .87]	.98 [.97, .99]	.98 [.98, .99]
C-2_A	6.90 ± 0.02 (<.001)	0.00 ± 0.01 (.99)	– 0.15 ± 0.01 (<.001)	– 0.01 ± 0.03 (.76)	0.15	1.01 [0.92, 1.1]	.73 [.62, .82]	.98 [.97, .99]	.98 [.97, .99]
C-3_A	0.07 ± 0.01 (<.001)	– 0.01 ± 0.01 (.14)	– 0.01 ± 0.01 (.05)	– 0.01 ± 0.01 (.14)	0.06	1.19 [0.7, 2.37]	.37 [.11, .72]	.91 [.62, .96]	.92 [.80, .96]
C-4_A	0.09 ± 0.01 (<.001)	0.00 ± 0.01 (.48)	– 0.01 ± 0.01 (.24)	– 0.01 ± 0.01 (.53)	0.07	1.13 [0.84, 1.55]	.47 [.24, .70]	.85 [.72, .91]	.93 [.88, .96]
C-5_A	5.82 ± 0.02 (<.001)	– 0.02 ± 0.01 (.03)	– 0.07 ± 0.01 (<.001)	– 0.01 ± 0.02 (.66)	0.13	1.07 [0.91, 1.24]	.66 [.56, .74]	1.00 [.99, 1.00]	1.00 [1.00, 1.00]
C-6_A	0.21 ± 0.01 (<.001)	0.01 ± 0.01 (.15)	0.02 ± 0.01 (<.01)	– 0.01 ± 0.01 (.38)	0.07	1.16 [0.97, 1.38]	.47 [.29, .62]	.69 [.54, .79]	.70 [.55, .81]
C-7_A	0.00 ± 0.00 (<.001)	0.00 ± 0.00 (.19)	0.00 ± 0.00 (1.00)	0.00 ± 0.00 (.50)	0.00	1.35 [1.07, 1.67]	.45 [.23, .62]	.57 [.31, .74]	.58 [.36, .73]
C-8_A	0.19 ± 0.02 (<.001)	– 0.01 ± 0.01 (.44)	0.03 ± 0.01 (<.01)	– 0.01 ± 0.02 (.79)	0.13	1.18 [1.01, 1.38]	.61 [.49, .71]	.85 [.78, .90]	.79 [.69, .87]
C-9_A	6.74 ± 0.03 (<.001)	0.01 ± 0.02 (.34)	– 0.21 ± 0.02 (<.001)	0.02 ± 0.04 (.52)	0.21	1.01 [0.87, 1.2]	.70 [.59, .80]	.99 [.98, .99]	.98 [.97, .99]
C-10_A	– 0.12 ± 0.02 (<.001)	0.00 ± 0.01 (.65)	0.02 ± 0.01 (.14)	0.01 ± 0.02 (.59)	0.13	0.93 [0.79, 1.1]	.57 [.43, .71]	.80 [.72, .86]	.72 [.58, .81]
C-11_A	0.07 ± 0.01 (<.001)	0.00 ± 0.00 (.79)	– 0.02 ± 0.00 (<.001)	– 0.01 ± 0.01 (.21)	0.00	1.11 [0.86, 1.45]	.67 [.47, .82]	.87 [.77, .92]	.93 [.87, .96]
C-12_A	0.07 ± 0.01 (<.001)	0.01 ± 0.01 (.52)	– 0.01 ± 0.01 (.23)	0.02 ± 0.02 (.33)	0.10	1.02 [0.82, 1.3]	.45 [.18, .67]	.72 [.49, .84]	.76 [.58, .85]
Measures from diffusion modeling									
DM-1_A	2.02 ± 0.06 (<.001)	0.01 ± 0.04 (.73)	– 0.22 ± 0.04 (<.001)	– 0.04 ± 0.07 (.60)	0.04	0.97 [0.83, 1.14]	.53 [.41, .64]	.72 [.59, .81]	.74 [.64, .81]
DM-2_A	0.51 ± 0.02 (<.001)	– 0.01 ± 0.02 (.64)	– 0.01 ± 0.02 (.61)	– 0.02 ± 0.02 (.29)	0.02	1.04 [0.9, 1.19]	.19 [.02, .36]	.60 [.44, .72]	.55 [.34, .69]
DM-3_A	0.36 ± 0.01 (<.001)	– 0.01 ± 0.01 (.20)	– 0.02 ± 0.01 (.01)	– 0.01 ± 0.01 (.50)	0.01	0.98 [0.81, 1.19]	.37 [.20, .53]	.49 [.29, .63]	.46 [.22, .63]
DM-4_A	1.72 ± 0.10 (<.001)	0.09 ± 0.07 (.17)	0.36 ± 0.07 (<.001)	0.18 ± 0.12 (.12)	0.07	0.96 [0.82, 1.13]	.48 [.34, .60]	.79 [.66, .87]	.79 [.70, .86]
DM-5_A	1.99 ± 0.06 (<.001)	– 0.04 ± 0.05 (.38)	– 0.23 ± 0.05 (<.001)	0.04 ± 0.07 (.61)	0.05	0.98 [0.74, 1.35]	.40 [.27, .53]	.65 [.53, .75]	.79 [.70, .86]
DM-6_A	0.34 ± 0.02 (<.001)	– 0.01 ± 0.01 (.61)	0.02 ± 0.01 (.06)	0.03 ± 0.02 (.05)	0.01	1.03 [0.88, 1.22]	.21 [– .01, .41]	.47 [.20, .64]	.52 [.31, .67]
DM-7_A	0.51 ± 0.01 (<.001)	0.01 ± 0.01 (.48)	– 0.03 ± 0.01 (.02)	0.01 ± 0.02 (.46)	0.01	0.76 [0.62, 0.93]	.43 [.25, .58]	.66 [.52, .77]	.80 [.71, .87]
DM-8_A	1.98 ± 0.11 (<.001)	0.03 ± 0.07 (.66)	0.24 ± 0.07 (<.001)	0.08 ± 0.11 (.45)	0.07	1.08 [0.86, 1.36]	.39 [.24, .55]	.80 [.72, .87]	.84 [.76, .90]
DM-09_A	5.04 ± 0.14 (<.001)	0.23 ± 0.09 (.02)	0.28 ± 0.09 (<.01)	– 0.01 ± 0.18 (.98)	1.20	0.86 [0.74, 1.01]	.60 [.49, .68]	.88 [.83, .92]	.89 [.84, .93]
DM-10_A	– 2.85 ± 0.25 (<.001)	– 0.08 ± 0.17 (.63)	0.27 ± 0.17 (.12)	– 0.09 ± 0.27 (.74)	1.74	1.02 [0.84, 1.25]	.43 [.26, .60]	.76 [.68, .82]	.80 [.73, .87]
DM-11_A	– 1.11 ± 0.15 (<.001)	0.18 ± 0.08 (.02)	– 0.47 ± 0.08 (<.001)	– 0.08 ± 0.16 (.61)	0.87	1.23 [1.04, 1.43]	.66 [.51, .78]	.77 [.68, .85]	.85 [.78, .90]
DM-12_A	2.01 ± 0.12 (<.001)	0.02 ± 0.07 (.78)	0.57 ± 0.07 (<.001)	0.12 ± 0.13 (.39)	0.80	1.07 [0.88, 1.27]	.57 [.43, .68]	.83 [.77, .88]	.90 [.84, .94]

CI = confidence interval; p = p-value; r_{Home} = internal reliability at home; r_{Lab} = internal reliability in the lab; r_{TR} = test–retest stability (lag: seven days), β_0 = intercept; β_1 = setting (home = 0, lab = 1); β_2 = number of session (1st session = 0, 2nd session = 1); β_3 = initial setting (home = 0, lab = 1); σ = standard deviation (*SD*) in the lab; ω = *SD* ratio with σ as reference; *SE* = standard error; labeling according to Table 4

References

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2015). QRTEngine: An easy solution for running online reaction time experiments using Qualtrics. *Behavior Research Methods*, *47*, 918–929. doi:https://doi.org/10.3758/s13428-014-0530-7

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, *57*, 289–300.

Bodnar, O., Link, A., Arendacká, B., Possolo, A., & Elster, C. (2017). Bayesian estimation in random effects meta-analysis using a non-informative prior. *Statistics in Medicine*, *36*, 378–399.

Brand, A., & Bradley, M. T. (2012). Assessing the effects of technical variance on the statistical outcomes of Web experiments measuring response times. *Social Science Computer Review*, *30*, 350–357.

Chetverikov, A., & Upravitelev, P. (2016). Online versus offline: The Web as a medium for response time data collection. *Behavior Research Methods*, *48*, 1086–1099. doi:https://doi.org/10.3758/s13428-015-0632-x

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*, e57410.

Davidson, D. J., Zacks, R. T., & Williams, C. C. (2003). Stroop interference, practice, and aging. *Aging, Neuropsychology, and Cognition: Section B*, *10*, 85–98. doi:https://doi.org/10.1076/anec.10.2.85.14463

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*, *48*, 1–12. doi:https://doi.org/10.3758/s13428-015-0567-2

Enge, S., Behnke, A., Fleischhauer, M., Küttler, L., Kliegel, M., & Strobel, A. (2014). No evidence for true training and transfer effects after inhibitory control training in young healthy adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 987–1001.

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General*, *137*, 201–225. doi:https://doi.org/10.1037/0096-3445.137.2.201

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*, 847–857.

Gomez, P., Ratcliff, R., & Perea, M. (2007). A model of the go/no-go task. *Journal of Experimental Psychology: General*, *136*, 389–413. doi:https://doi.org/10.1037/0096-3445.136.3.389

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust Web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*, 93–104.

Hecht, H., Oesker, M., Kaiser, A., Civelek, H., & Stecker, T. (1999). A perception experiment with time-critical graphics animation on the World-Wide Web. *Behavior Research Methods, Instruments, & Computers*, *31*, 439–445.

Hilbig, B. E. (2016). Reaction time effects in lab- versus Web-based research: Experimental evidence. *Behavior Research Methods*, *48*, 1718–1724. doi:https://doi.org/10.3758/s13428-015-0678-9

Keller, F., Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing accuracy of Web experiments: A case study using the WebExp software package. *Behavior Research Methods*, *41*, 1–12. doi:https://doi.org/10.3758/BRM.41.1.12

Lerche, V., & Voss, A. (2017). Retest reliability of the parameters of the Ratcliff diffusion model. *Psychological Research*, *81*, 629–652.

Miller, J., & Ulrich, R. (2013). Mental chronometry and individual differences: Modeling reliabilities and correlations of reaction time means and effect sizes. *Psychonomic Bulletin & Review*, *20*, 819–858. doi:https://doi.org/10.3758/s13423-013-0404-5

Miller, R., Scherbaum, S., Heck, D. W., Goschke, T., & Enge, S. (2017). On the relation between the (censored) shifted Wald and the Wiener distribution as measurement models for choice response times. *Applied Psychological Measurement*, *42*, 116–135. doi:https://doi.org/10.1177/0146621617710465

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*, 49–100. doi:https://doi.org/10.1006/cogp.1999.0734

Neath, I., Earle, A., Hallett, D., & Surprenant, A. M. (2011). Response time accuracy in Apple Macintosh computers. *Behavior Research Methods*, *43*, 353–362. doi:https://doi.org/10.3758/s13428-011-0069-9

Paap, K. R., & Sawi, O. (2016). The role of test–retest reliability in measuring individual and group differences in executive functioning. *Journal of Neuroscience Methods*, *274*, 81–93.

Plant, R. R., & Quinlan, P. T. (2013). Could millisecond timing errors in commonly used equipment be a cause of replication failure in some neuroscience studies? *Cognitive, Affective, & Behavioral Neuroscience*, *13*, 598–614. doi:https://doi.org/10.3758/s13415-013-0166-6

Plant, R. R., & Turner, G. (2009). Millisecond precision psychological research in a world of commodity computers: New hardware, new problems? *Behavior Research Methods*, *41*, 598–614. doi:https://doi.org/10.3758/BRM.41.3.598

R Core Team. (2017). R: A language and environment for statistical computing (Version 3.3.1). Vienna: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Reimers, S., & Maylor, E. A. (2005). Task switching across the life span: Effects of age on general and specific switch costs. *Developmental Psychology*, *41*, 661–671. doi:https://doi.org/10.1037/0012-1649.41.4.661

Reimers, S., & Stewart, N. (2007). Adobe Flash as a medium for online experimentation: A test of reaction time measurement capabilities. *Behavior Research Methods*, *39*, 365–370. doi:https://doi.org/10.3758/BF03193004

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods*, *47*, 309–327. doi:https://doi.org/10.3758/s13428-014-0471-1

Reips, U.-D. (2002). Standards for internet-based experimenting. *Experimental Psychology*, *49*, 243–256. doi:https://doi.org/10.1027/1618-3169.49.4.243

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*, 207–231. doi:https://doi.org/10.1037/0096-3445.124.2.207

Schmitz, F., & Voss, A. (2012). Decomposing task-switching costs with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 222–250.

Schubert, T. W., Murteira, C., Collins, E. C., & Lopes, D. (2013). ScriptingRT: A software library for collecting response latencies in online studies of cognition. *PLoS ONE*, *8*, e67769. doi:https://doi.org/10.1371/journal.pone.0067769

Semmelmann, K., & Weigelt, S. (2017). Online psychophysics: Reaction time effects in cognitive experiments. *Behavior Research Methods*, *49*, 1241–1260.

Simcox, T., & Fiez, J. A. (2014). Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behavior Research Methods*, *46*, 95–111. doi:https://doi.org/10.3758/s13428-013-0345-y

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, *21*, 736–748. doi:https://doi.org/10.1016/j.tics.2017.06.007

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology*, *60*, 385–402.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*, 1206–1220. doi:https://doi.org/10.3758/BF03196893

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, *39*, 767–775. doi:https://doi.org/10.3758/BF03192967

Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*, 641–671. doi:https://doi.org/10.1080/09541440802205067

Willoughby, M., & Blair, C. (2011). Test–retest reliability of a new executive function battery for use in early childhood. *Child Neuropsychology*, *17*, 564–579. doi:https://doi.org/10.1080/09297049.2011.554390

Wolff, M., Krönke, K.-M., Venz, J., Kräplin, A., Bühringer, G., Smolka, M. N., & Goschke, T. (2016). Action versus state orientation moderates the impact of executive functioning on real-life self-control. *Journal of Experimental Psychology: General*, *145*, 1635–1653.