Atten Percept Psychophys (2014) 76:2495–2507
DOI 10.3758/s13414-014-0745-9

Variability in color-choice Stroop performance within and across EEG and MRI laboratory contexts

T. Fehr · J. Wiechert · P. Erhard

Published online: 2 August 2014

Abstract To examine the reproducibility of the Stroop effect, behavioral data from 22 healthy female individuals were repeatedly (three-month interval between two separate measurement sessions) obtained while performing a color-choice Stroop task under realistic magnetic resonance imaging (MRI) and electroencephalography (EEG) laboratory conditions. At the group statistical level, the Stroop effect, indicated by longer response times for incongruent than for congruent stimulus conditions, was consistently present for almost all examined measurement levels. However, differential effects of laboratory contexts on retest reproducibility were observed across repeated measurement levels, both within and between sessions. These results challenge existing theories about the underlying nature of Stroop interference processing. It appears necessary to apply a multitheoretical approach, because intraindividual variability within and across measurement sessions suggests potential fluctuations in the individual mental strategies applied, recruitment of varying memory resources, the influence of mediator variables such as working memory capacity and/or attention, and many more possible variations. Single-observation studies run the risk of favoring a single theoretical concept and therefore underestimating the individual factor. We further conclude that dependent analysis-of-variance statistics are a more fit test for reproducibility than are correlative reliability estimations.

Keywords Stroop interference · Cognition · Repeated measures · Experimental context · Reliability

The reliability and reproducibility of experimental data have frequently been postulated as being among the key features of clinical and scientific procedures (e.g., Siegrist, 1995, 1997; Strauss, Allen, Jorgensen, & Cramer, 2005; Wöstmann et al., 2013). These data requirements are even more essential in psychophysiological approaches that claim to identify the involvement of brain structures in both simple and complex mental processing (e.g., Button et al., 2013; Fehr, 2013; Thirion et al., 2007). Reliability, however, is based on a correlative parameter that refers to consistency in the rank order among individuals across measurement sessions. In the present work, reliability coefficients are reported, but the main focus is laid on Stroop-effect sizes related to the consistency of intra- and inter-measurement-session and inter-methodological-context magnitudes, examined by means of repeated measures analysis-of-variance (ANOVA) approaches.

Inconsistencies in results between studies or repeated measurement sessions that have putatively examined the same or similar complex mental processes (e.g., Achtziger, Fehr, Oettingen, Gollwitzer, & Rockstroh, 2008; Fehr, 2013; Fehr, Achtziger, Hinrichs, & Herrmann, 2003; Thirion et al., 2007),

T. Fehr (✉) · P. Erhard
Center for Cognitive Sciences, University of Bremen, Hochschulring 18, 28359 Bremen, Germany
e-mail: fehr@uni-bremen.de

P. Erhard
e-mail: erhard@uni-bremen.de

T. Fehr
Department of Neuropsychology, University of Bremen, Bremen, Germany

T. Fehr · P. Erhard
Center for Advanced Imaging, Universities of Bremen and Magdeburg, Bremen, Germany

J. Wiechert
Department of Neurology, University of Lübeck, 23562 Lübeck, Germany
e-mail: juliana.wiechert@neuro.uni-luebeck.de

P. Erhard
AG In Vivo MR, University of Bremen, Bremen, Germany

but also basic psychophysical processes (e.g., Fahle & Spang, 2003), have been discussed as an important issue. Despite its significance, however, both inter- and intraindividual variability are often neglected in clinical and experimental neuroscientific group studies. It has also been shown that laboratory environments, such as a noise-generating magnetic resonance imaging (MRI) scanner, selectively influence performance in a complex working memory task (Tomasi, Caparelli, Chang, & Ernst, 2005).

One experimental approach that has provided highly robust behavioral effects is the classic Stroop task, first introduced by John Ridley Stroop in 1935. Text-color naming of colored color words, whose visual color does not match the word meaning (incongruent task), have consistently produced longer response or reading times and, dependent on the experimental design, more errors than text-color naming of colored color words whose visual color matches the word meaning (congruent task). This task has been modified and successfully applied in many different versions in clinical and scientific contexts (e.g., MacLeod, 1998, 2005; MacLeod & MacDonald, 2000).

However, different studies have reported a large variety of brain structures as being potentially involved in Stroop task processing. Thus, despite numerous sound theoretical concepts explaining the Stroop effect (e.g., Kane & Engle, 2003; MacLeod, 2005; Melara & Algom, 2003; Roelofs, 2003, 2010; Unsworth & Engle, 2007), key problems remain in verifying or falsifying the different proposed theoretical models. For example, one of the most severe problems is that congruent and incongruent task performance has been shown to be reliable, but if the parameters are combined in mean difference values (known as the "Stroop effect"), a more complex parameter results, which has been shown to run the risk of lacking reliability (e.g., Siegrist, 1997; Strauss et al., 2005).

At the same time, however, the Stroop effect cannot be completely erased at the ANOVA-related group statistical level (cf. Ackerman & Schneider, 1984; Beglinger et al., 2005; Davidson, Zacks, & Williams, 2003; Ellis & Dulaney, 1991). The reason for this might be that variations in the rank orders of performance among individuals across measurement sessions produced reduced reliability estimations, but no actual change in individual mean differences in performance measurement for incongruent versus congruent task performance.

Several basic hypotheses were investigated in the present study:

1. The Stroop effect was expected to be always present at an ANOVA-related group statistical level for different time segments within the experimental run and for all repeated measurement levels (Ackerman & Schneider, 1984; Beglinger et al., 2005; Davidson et al., 2003; Ellis & Dulaney, 1991).
2. Since learning effects have been reported for different Stroop-task designs (e.g., Ellis & Dulaney, 1991; MacLeod, 1998), performance was expected to improve over time within the experimental runs as well as over repeated measurement sessions.
3. If scanner noise in the MRI environment selectively affects more complex tasks, as has been demonstrated for tasks with higher working memory load by Tomasi et al. (2005), then Stroop performance should be reduced in particular for incongruent tasks in the MRI, but not in the electroencephalography (EEG), environment.
4. We additionally assumed that reliability (i.e., the consistency of rank orders among study participants across repeated measurements) would be less present for Stroop-effect data than for response time data of the incongruent and congruent stimulus conditions considered separately (e.g., Siegrist, 1997; Strauss et al., 2005).

Method

Study participants

In the present study, 24 healthy, right-handed female psychology students at the University of Bremen, Germany, were included. Two participants were excluded from the sample because they did not participate in all of the measurement sessions. This focus on one gender group should exclude potential gender effects. Participants received a certificate for their participation; these certificates are needed for achieving a psychology degree at German universities. Students freely chose the studies in which they wanted to participate. The remaining 22 participants ranged from 21 to 29 years of age (mean age: 23.8 ± 1.8 years), and did not report any psychological or neurological disorders. After the first measurement session, participants filled in the Mehrfach-Wahl-Wortschatz-Test, B-version (MWT-B; Lehrl, 2005), a German short test to determine general intelligence level.

All participants were thoroughly introduced to the experimental environment and confirmed their participation by written consent. The study was approved by the local ethics committee.

Experimental design

In the present report, we provide part of the data acquired in a larger study protocol; the presented data were always obtained in the first of two measurement runs in each measurement session. The stimulus sequence (see the next paragraph for details) was separately presented in EEG and MRI laboratory environments during measurement sessions. In two measurement sessions (s1 and, about three months later, s2), the stimuli were presented to ten individuals, first under EEG, and then, one to three days later, under MRI laboratory conditions. The stimuli were first presented to the remaining 12 participants under MRI and then under EEG conditions.

The MRI laboratory environment was different from the EEG laboratory environment in several ways: MRI measurement sessions were characterized by scanner noise (echo planar imaging; ear plugs with a >25-dB attenuation were used by all participants), and participants were in a lying-down position in a narrow scanner tube with their head embedded in foam material to minimize movement. The EEG measurement session was characterized by silence, sitting upright, and a tight EEG cap on participants' heads.

Each trial of the Stroop task (cf. Fehr, Wiedenmann, & Herrmann, 2006) consisted of a colored letter sequence and four colored rectangles placed below (red, blue, green, and yellow, shuffled from trial to trial in order to avoid a color–motor-response association; see also below) located in the center of the respective display facility (a flat-screen monitor in the case of EEG measurement, and a display on a back-projection screen observed via a mirror in the MRI scanner). In both measurement environments, the visual angles of the displayed stimuli were the same and were kept below 3.8 deg in both the horizontal and vertical directions (please see Fig. 1 for details of the trial elements and trial timing).

Three different stimulus conditions were presented in the Stroop design: (1) In a baseline condition (bas), a sequence of colored Xs ("XXXX") was displayed; (2) in the congruent condition (con), a color word (e.g., "green") was presented in the color that matched the meaning of the word (here, green); and (3) in the incongruent condition (inc), a color word was presented in a color that did not match the meaning of the word (e.g., the word "red," printed in either blue, green, or yellow; see Fig. 1 for a detailed illustration). Study participants were asked to identify the color of the displayed letter sequence while ignoring potential word meanings, and to press the button that corresponded to the colored rectangles below the letter sequence as quickly as possible. There was a forced choice time window of 1,500 ms, during which the participants were required to respond. If the participant needed longer than 1,500 ms to respond, the message *zu langsam!* ("too slow!") was displayed in the center of the display facility. If the participant responded wrongly, *falsch* ("wrong") was displayed, and in case of a correct response, *richtig* ("correct") was displayed. After feedback display (2,000 ms) a centered fixation dot was presented during a between-stimulus interval that was pseudorandomly jittered between 2,600 and 3,400 ms.

Incongruent, congruent, and baseline stimulus trials were presented in a pseudorandomized sequential order (96 trials per stimulus condition). No stimulus type was consecutively repeated more than three times. The sequence of the different stimulus types was balanced across the trials (each sequential combination—such as, e.g., "inc after con," "con after inc," etc.—occurred 32 times, except for "inc after bas," which occurred 31 times). Before every measurement session, participants were presented with 15 practice trials to familiarize them with the task and the response buttons.

Statistical analyses

Several different analysis strategies were applied to the existing data. The behavioral data originated from four runs of the same Stroop task, two of them in an MRI environment and two of them in an EEG environment (see above). The statistical analyses included the analysis of changes in the behavioral data over time segments, over sessions, between methodological laboratory contexts, and also a within-session split-half analysis.

The reported statistical analyses were related to the response times in congruent and incongruent stimulus conditions separately, and to the mean difference values of both parameters. To avoid potential influences of short-term habituation or initial practice effects within the measurement runs, the first 12 trials of each stimulus condition (36 trials in all) were omitted from all analyses.

The remaining 84 trials per stimulus condition were statistically processed as follows: For intrasession parameter estimations, even and odd trials within one stimulus condition were pooled in two separate levels (r1 and r2) of an intrasession split-half factor (for a detailed illustration, see Fig. 1B). This approach of interleaved split-half trial splitting was used to avoid time segment effects, such as learning effects or tiredness or vigilance fluctuations. Following this analysis, two statistical intrasession repeated measurement cells (r1 and r2) were filled for each stimulus condition (bas, con, and inc) in each repeated measurement session (intersession: s1 and s2) and each measurement modality (MRI and EEG environments; see Figs. 5 and 6 below for a schematic illustration of the repeated measurement design and results).

Following the above-described procedures, for the con and inc stimulus conditions, the individual mean response times of two repeated measurement levels for each EEG and MRI session were calculated separately. Thus, intra- and intersession differences could be estimated separately for both the EEG and MRI conditions (intrasession interleaved split-half: s1r1 vs. s1r2 and s2r1 vs. s2r2; and intersession: s1r1 vs. s2r1 and s1r2 vs. s2r2). Intercontextual variability was also proven for each repeated measurement factor separately: EEG–s1r1 versus MRI–s1r1, EEG–s1r2 versus MRI–s1r2, EEG–s2r1 versus MRI–s2r1, and EEG–s2r1 versus MRI–s2r1.

In addition, to examine potential intrasession time-segment effects, the response time data were pooled for the first,

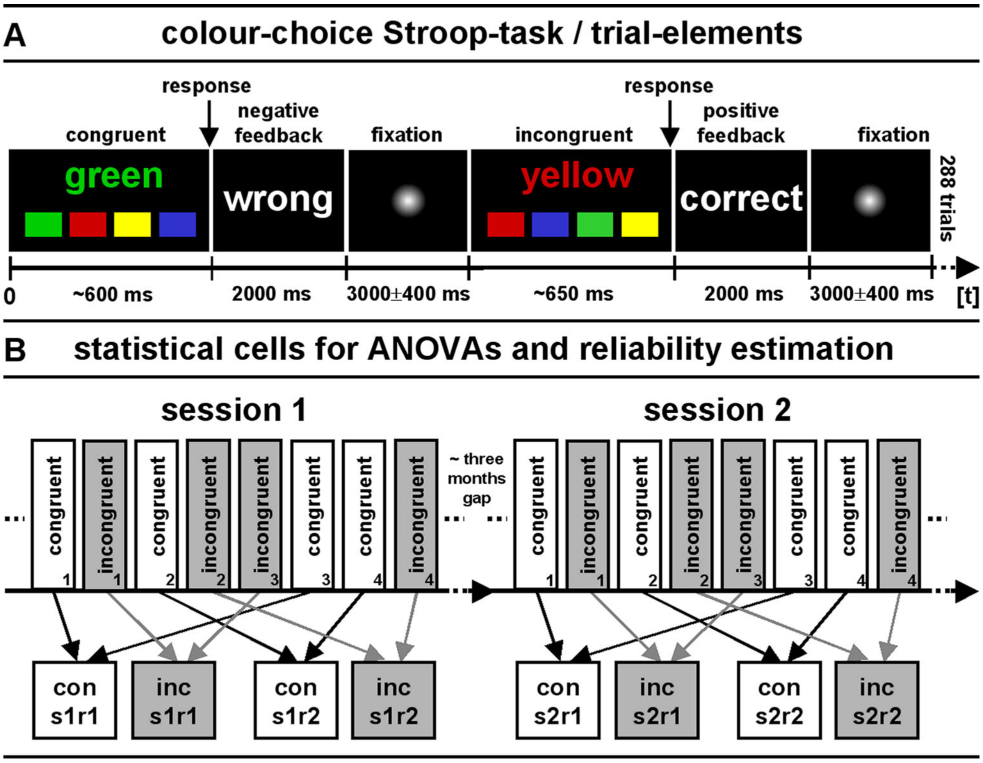


Fig. 1 (A) Illustration of the experimental design, trial timing, and trial elements. After task presentation in each trial, the participants were asked to press one of four buttons (left and right index and middle fingers, placed either on the keys "D," "F," "J," and "K" at the computer keyboard, in the case of the EEG conditions, or on the respective buttons on two two-button devices, specially constructed for MRI facilities, held in the left and right hands), corresponding to which color on the bar placed below matched the color of the corresponding colored letter string above. (B) Statistical cells for both the ANOVA and reliability estimations (the fMRI and EEG sessions followed the same statistical design schema). Within each session (s1 and s2), all even and odd trials for each stimulus condition were pooled separately into so-called "interleaved split-half cells" (r1 and r2)

second, and third time segments of trials in the whole sequence (resulting in 28 trials for each stimulus condition and time segment). A repeated measures ANOVA including the factors Time Segment (three levels: 1st, 2nd, and 3rd), Method, Session, and Stimulus Condition was applied. Additionally, a repeated measures ANOVA was calculated including the factors Time Segment, Method, and Session on the Stroop-effect data. Subsequent *t* tests on repeated measurement data and statistical tests against zero were additionally performed.

To test for general quantitative differences in response times between methods, sessions, and stimulus conditions, a repeated measures ANOVA including the factors Method (two levels: EEG and fMRI), Session (two levels: initial session and follow-up session about three months later), and Stimulus Condition (two levels: con and inc stimuli) was calculated. Separately, a repeated measures ANOVA including the factors Method and Session was applied to the Stroop-effect response time data. Subsequently, post-hoc paired *t* test estimations were performed to test for differences between the respective repeated measurement factors.

As a measure of consistency of rank orders among the study participants across repeated measurement levels (cf. Golden, Sawicki, & Franzen, 1984), reliability was explored by calculating Spearman–Brown reliability coefficients (calculated as $r - 1 / r * 2$, where r = Pearson correlation coefficient).

Results

Error rates in the con and inc stimulus conditions were on average below 0.015% ± 0.002% and 0.014% ± 0.003% across measurement sessions and individuals, respectively, and therefore will not be presented or discussed in the following sections. MWT-B, as a fast measure of general intelligence (Lehrl, 2005), obtained after the first physiological measurement session (EEG or fMRI), showed IQ values between 92 and 136 points (103.8 ± 13.9), which documents an average intelligence-level distribution across participants.

Response time differences between the inc and con task conditions

Response time data and statistical effects are illustrated in Fig. 2A for the con and inc stimulus conditions separately, and in Fig. 2B for the Stroop-effect data. We found a

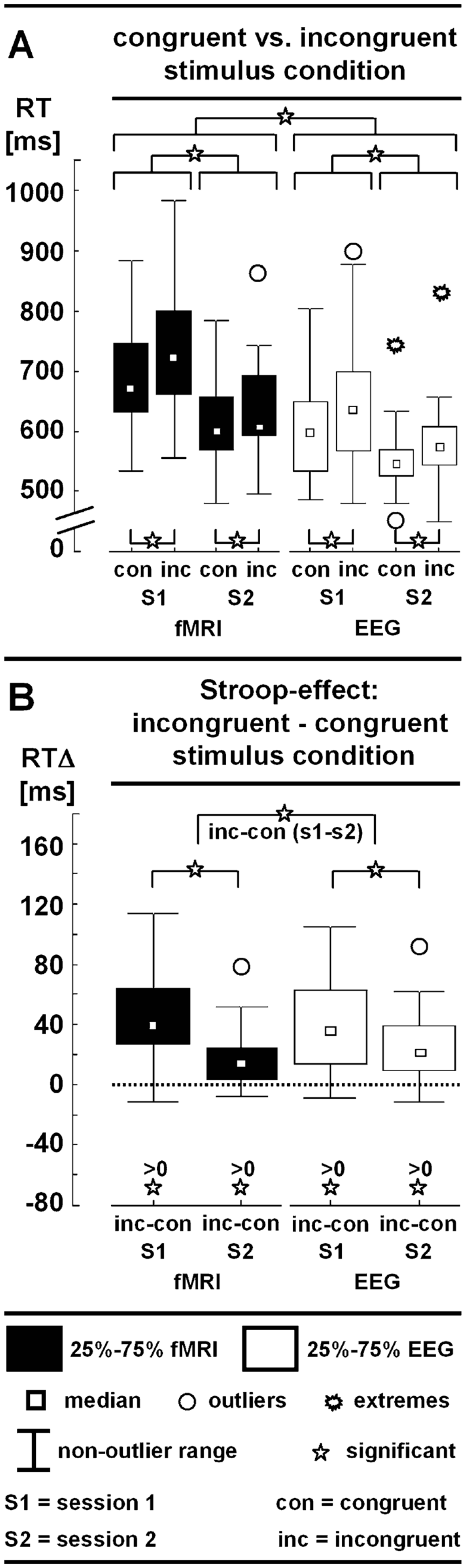


Fig. 2 Boxplot diagrams of the response time data from each repeated ▶ measurement session (s1 and s2) and measurement context (MRI and EEG), and an illustration of the statistical test results, for the congruent and incongruent task conditions separately (A) and for difference values (=the Stroop effect) between the two conditions (B)

significant Method × Session × Stimulus Condition effect [*F*(1, 21) = 5.4, $p < .05$; Greenhouse–Geisser adjustment (GG): $p < .05$], a Session × Condition effect [*F*(1, 21) = 32.2, $p < .05$; GG: $p < .05$], and a Method × Session effect [*F*(1, 21) = 12.3, $p < .05$; GG: $p < .05$]. Subsequent post-hoc analyses (paired *t* tests, where $p < .05$ was considered significant; see Fig. 2A) indicated that for all intra- and intersession repeated measurement factors, inc tasks produced longer response times than did con tasks [stimulus condition effect: *F*(1, 21) = 42.1, $p < .05$; GG: $p < .05$]. Response times were generally longer in the MRI than in the EEG environment across all stimulus conditions [method effect: *F*(1, 21) = 43.0, $p < .05$; GG: $p < .05$]. Furthermore, response times were generally faster in the second than in the first measurement session in both the MRI and EEG environments [session effect: *F*(1, 21) = 57.0, $p < .05$; GG: $p < .05$].

The Stroop effect showed a significant Method × Session interaction [*F*(1, 21) = 5.4, $p < .05$, GG: $p < .05$] and a significant session effect [*F*(1, 21) = 32.1, $p < .05$; GG: $p < .05$]. All session- and method-related mean Stroop-effect response times differed significantly from zero (see Fig. 2B). In both measurement environments, MRI and EEG, the second measurement session yielded smaller Stroop effects. Differences between the Stroop-effect response times across sessions also differed between methods, indicating a larger session effect in the MRI than in the EEG environment (paired *t* test: $t = 2.3$, $p < .05$).

Intrasession time-segment effects in Stroop performance

Intrasession time-segment effects were examined on pooled data in three consecutive time segments of 28 trials each. Figure 3A illustrates comparisons of the response times between stimulus conditions within each of the three time segments. Figure 3B illustrates the respective Stroop effects in each time segment, and Fig. 4 shows time-segment-related time courses of response times in each stimulus condition. Several subsequent time-segment-related factorial effects were found when the response time values of the congruent and incongruent stimulus conditions were separately included in repeated measures ANOVAs: Time Segment × Method × Session × Stimulus Condition, Time Segment × Condition [*F*(4, 84) = 7.1, $p < .05$; GG: $p < .05$], Time Segment × Session [*F*(2, 42) = 5.1, $p < .05$; GG: $p < .05$], and time segment [*F*(2, 42) = 17.5, $p < .05$; GG: $p < .05$]. Least significance difference (Fisher´s LSD) tests generally indicated longer

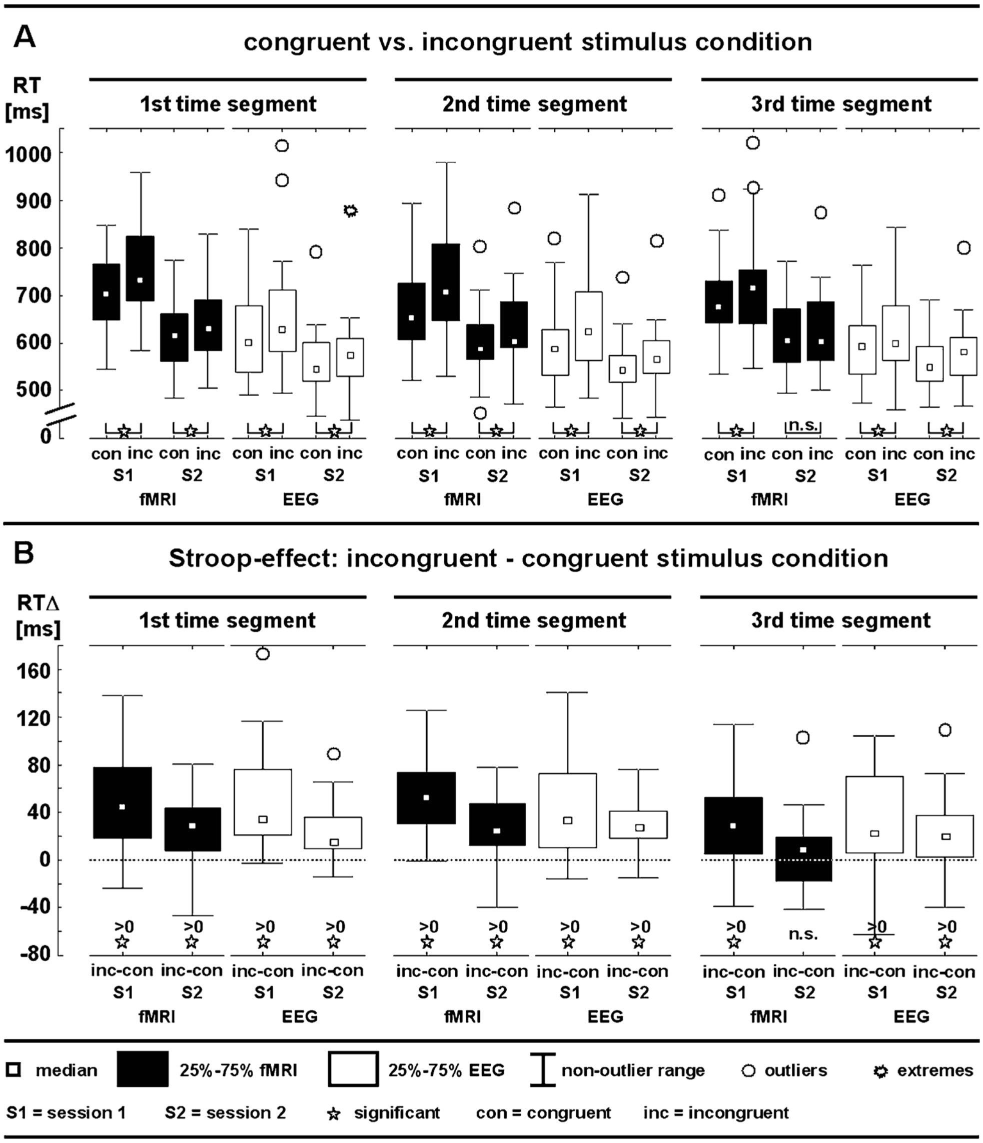


Fig. 3 Boxplot diagrams of the response time data of each repeated measurement session (s1 and s2) and measurement context (MRI and EEG), shown separately for three consecutive time segments across the experimental run. (A) Illustration of statistical test results for the congruent and incongruent task conditions separately (B) Illustration of statistical test results for difference values (=Stroop effect) between the two conditions. The figure was optimized to illustrate intersession and intermeasurement contextual (MRI and EEG) differences within the time segments

response times in the first time segment as compared to the second and third time segments (p < .05) in the first measurement session, whereas in the second measurement session, no general response time differences emerged across time segments.

Response times for the congruent condition showed larger values in the first and in the last time segment, as compared to the second time segment, forming a U-shaped time course (Fig. 4). Incongruent stimulus condition response times showed their largest values in the first time segment relative to both the second and third time segments of trials, forming a hyperbola-like time course (Fig. 4).

Response time differences between the incongruent and congruent stimulus conditions were apparent across all time segments, sessions, and methods (paired t tests, p < .05), except in the third time segment of MRI-context-related data in the second session.

The Stroop-effect data showed a significant time-segment effect [F(2, 42) = 7.8, p < .05; GG: p < .05], subsequently explained by a statistical trend (LSD, p < .1) demonstrating a smaller Stroop effect in the third time segment than in the first and second time segments. Across all time segments, sessions, and methods, the Stroop effects could be shown to be different from zero

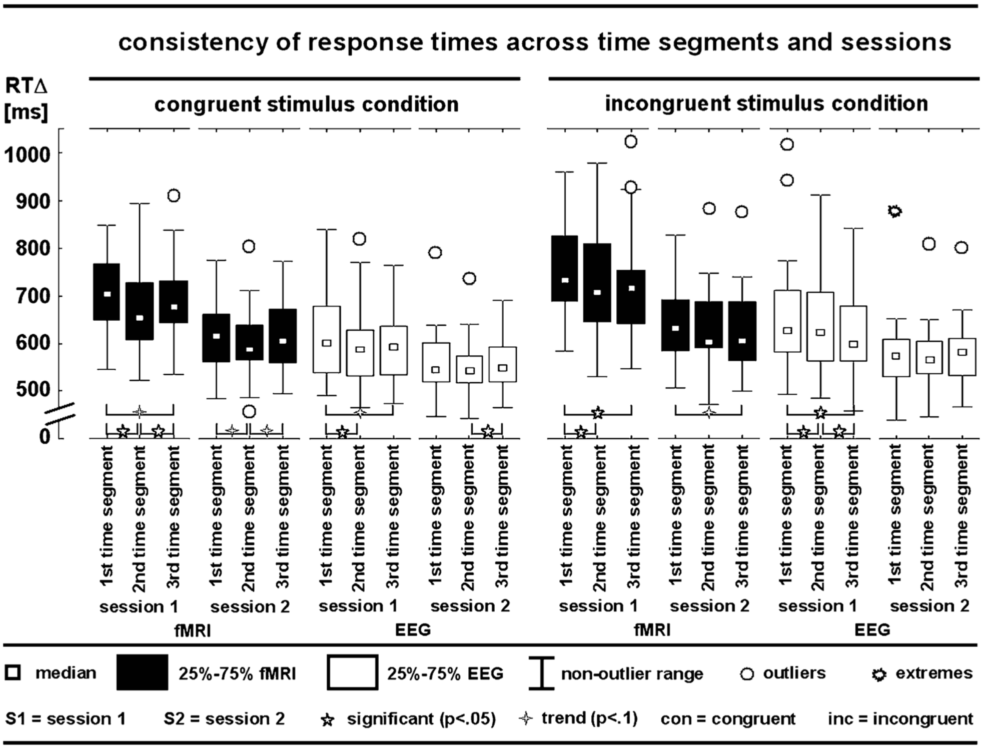


Fig. 4 Boxplot diagrams of the response time data of each repeated measurement session and measurement context (MRI and EEG), shown separately for three consecutive time segments across the experimental run: Illustration of statistical test results for the congruent and incongruent task conditions separately. The figure was optimized to illustrate the intra-task-condition dynamics within the respective experimental runs

(p < .05), except in the third time segment of MRI-context-related data in the second session.

Testing for quantitative intra- and intersession and intercontextual Stroop performance differences

To test the consistency of response time performance across the intra- and intersession data and across intercontextual repeated measurement levels, paired t tests and statistical power values were calculated. The detailed results are illustrated in Fig. 5A and B for the con and inc task conditions, respectively, and in Fig. 6 for the Stroop-effect data. Since we observed selective Stroop-effect differences between the MRI and EEG measurement environments in the third time segment of the presented task trials (see the previous section), statistical testing was restricted to the pooled data of the first two time segments of trials (56 trials for each stimulus condition, and 28 trials for each intrasession even- and odd-trial interleaved repeated measurement factor).

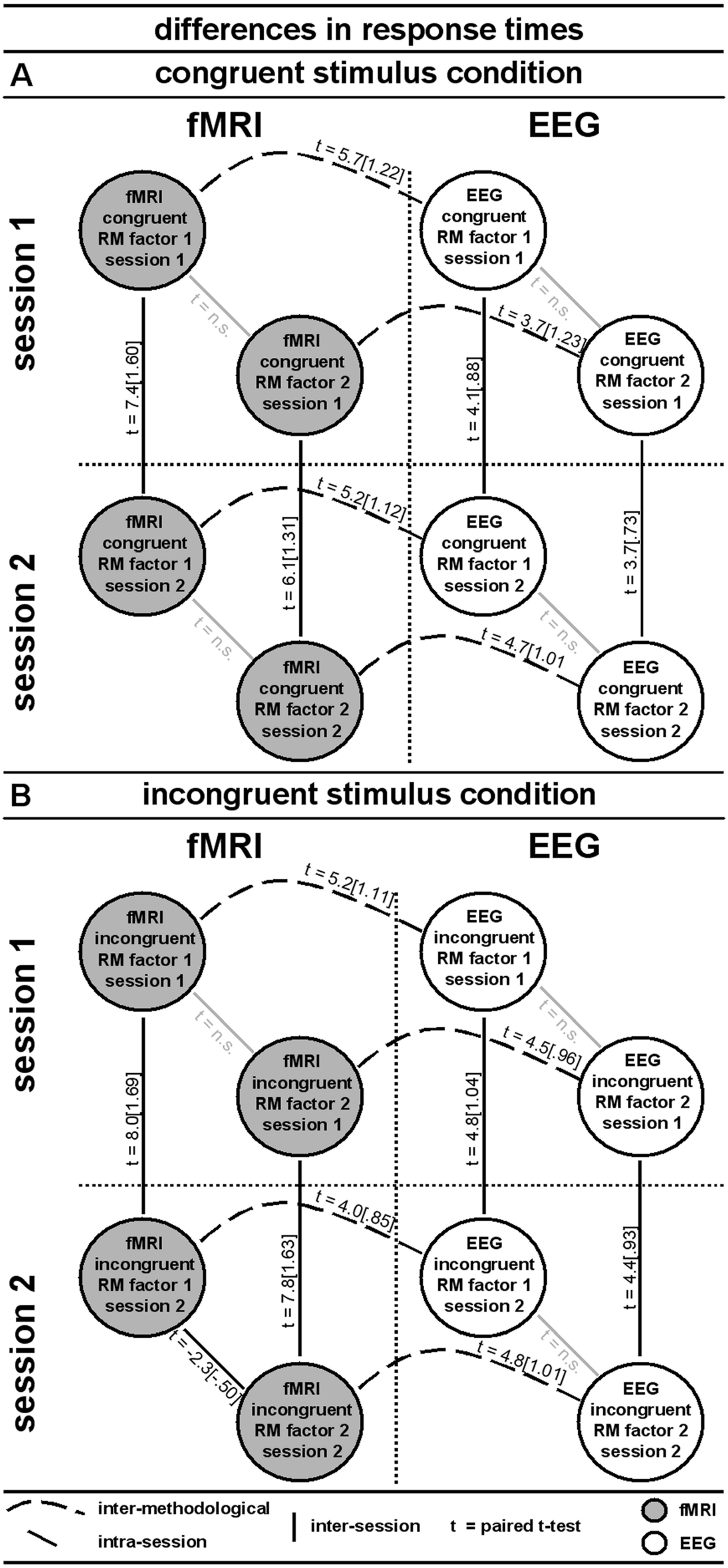
For both the con and inc task conditions, intersession and intercontextual context-related performance differences, explained by faster response times in the second than in the first session and by faster response times in the EEG than in the fMRI environment, could be shown. All of these differences were substantiated by higher levels of statistical power (see Fig. 5A and B for details). Only in the second session under MRI conditions did we find a significant performance difference between intrasession repeated measurement levels; this difference, however, was substantiated by moderate statistical power (see Fig. 5B for details).

For the Stroop-effect data, intrasession performance consistency could not be confirmed for the first measurement in the EEG environment and in the second measurement session in the fMRI environment (see Fig. 6 for details). Intersession differences could be explained by generally lower Stroop effects in the second measurement session (substantiated by average and high statistical power values; see Fig. 6 for details), whereas differences between the sessions were larger under MRI conditions. For both the first and the second measurement sessions, the intercontextual (MRI vs. EEG environment) consistency of the Stroop-effect data was confirmed by nonsignificant paired t tests.

Intra- and intersession and intercontextual reliability estimations

Reliability, as a measure of consistency in the rank orders among study participants across repeated measurement levels, was estimated by a Spearman–Brown reliability coefficient. The detailed results are listed in Table 1. Reliability estimates were restricted to the pooled data of the first two time segments of trials.

The examination of response times in the congruent and incongruent stimulus conditions showed separately that all


differences in response times
A
congruent stimulus condition
fMRI
EEG
session 1
session 2
fMRI congruent RM factor 1 session 1
fMRI congruent RM factor 2 session 1
fMRI congruent RM factor 1 session 2
fMRI congruent RM factor 2 session 2
EEG congruent RM factor 1 session 1
EEG congruent RM factor 2 session 1
EEG congruent RM factor 1 session 2
EEG congruent RM factor 2 session 2
t = 5.7[1.22]
t = n.s.
t = n.s.
t = 3.7[1.23]
t = 7.4[1.60]
t = 4.1[.88]
t = 5.2[1.12]
t = 6.1[1.31]
t = 3.7[.73]
t = n.s.
t = n.s.
t = 4.7[1.01]
B
incongruent stimulus condition
fMRI
EEG
session 1
session 2
fMRI incongruent RM factor 1 session 1
fMRI incongruent RM factor 2 session 1
fMRI incongruent RM factor 1 session 2
fMRI incongruent RM factor 2 session 2
EEG incongruent RM factor 1 session 1
EEG incongruent RM factor 2 session 1
EEG incongruent RM factor 1 session 2
EEG incongruent RM factor 2 session 2
t = 5.2[1.11]
t = n.s.
t = n.s.
t = 4.5[.96]
t = 8.0[1.69]
t = 4.8[1.04]
t = 4.0[.85]
t = 7.8[1.63]
t = 4.4[.93]
t = -2.3[-.50]
t = n.s.
t = 4.8[1.01]
inter-methodological
intra-session
inter-session
t = paired t-test
fMRI
EEG

◀ **Fig. 5** Illustration of statistical intra- and intersession and inter-methodological-context comparisons for response time data of the congruent and incongruent task conditions separately. For paired *t* tests, standardized Cohen's *d* values are added in brackets after the respective *t* values

intra- and intersession, as well as all intercontextual, reliability coefficients were based on significant correlations ($p < .05$; see Table 1 for details). This indicated that participants performed in comparable rank orders across all repeated measurement levels.

For the Stroop-effect data, reliability estimations provided a more complex picture: For the MRI context, reliability could be confirmed for neither the intra- nor the intersession repeated measurement levels. For the EEG context, all repeated measurement levels showed significant reliability estimations (underlying correlations, $p < .05$), except the intrasession levels in the second measurement session. The intercontextual reliability of the Stroop-effect data was significant for the first, but not for the second, measurement session (see Table 1 for details).

Discussion

The color-word Stroop interference task has been widely used in different scientific and clinical contexts. It has been modified and applied in many different ways. Most reported

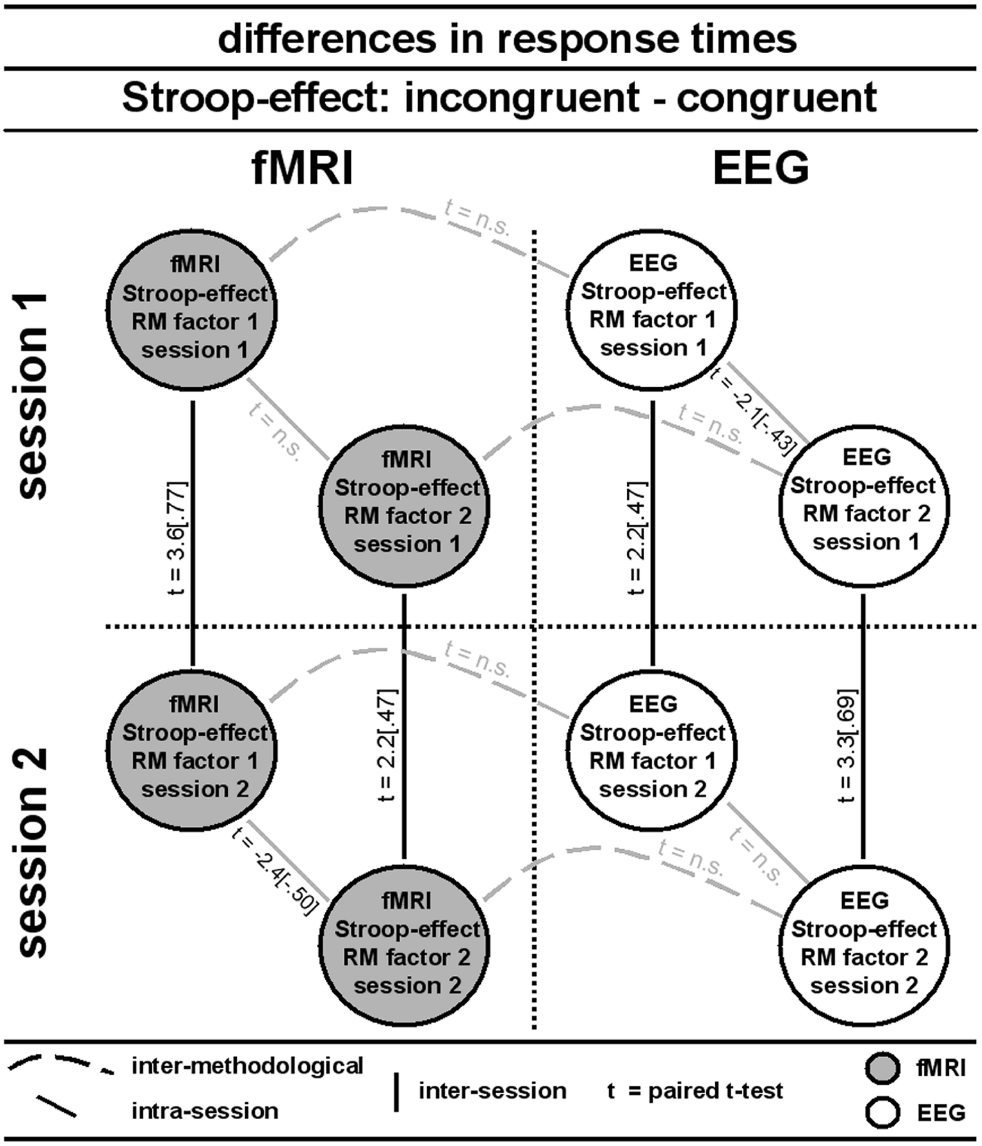

differences in response times
Stroop-effect: incongruent - congruent
fMRI
EEG
session 1
session 2
fMRI Stroop-effect RM factor 1 session 1
fMRI Stroop-effect RM factor 2 session 1
fMRI Stroop-effect RM factor 1 session 2
fMRI Stroop-effect RM factor 2 session 2
EEG Stroop-effect RM factor 1 session 1
EEG Stroop-effect RM factor 2 session 1
EEG Stroop-effect RM factor 1 session 2
EEG Stroop-effect RM factor 2 session 2
t = n.s.
t = n.s.
t = n.s.
t = -2.1[-.43]
t = 3.6[.77]
t = 2.2[.47]
t = n.s.
t = 2.2[.47]
t = 3.3[.69]
t = -2.4[-.50]
t = n.s.
t = n.s.
inter-methodological
intra-session
inter-session
t = paired t-test
fMRI
EEG

Fig. 6 Illustration of statistical intra- and intersession and inter-methodological-context comparisons for response time data of the congruent and incongruent task conditions separately. For paired *t* tests, standardized Cohen's *d* values are added in brackets after the respective *t* values

Table 1 Reliability coefficients (Spearman–Brown $R = [r \times 2] / [r + 1]$, where r = Pearsons´s correlation coefficient) for the different repeated measurement levels, indicating consistency of rank orders among study participants across repeated measurement levels

Repeated Measurement Levels	Reliability Coefficient
Reliability: Congruent Condition	
fMRI Context	
fMRIs1r1 ⊗ fMRIs1r2	.97
fMRIs2r1 ⊗ fMRIs2r2	.96
fMRIs1r1 ⊗ fMRIs2r1	.90
fMRIs1r2 ⊗ fMRIs2r2	.80
EEG Context	
EEGs1r1 ⊗ EEGs1r2	.99
EEGs2r1 ⊗ EEGs2r2	.96
EEGs1r1 ⊗ EEGs2r1	.88
EEGs1r2 ⊗ EEGs2r2	.84
Intercontextual: EEG/fMRI	
fMRIs1r1 ⊗ EEGs1r1	.78
fMRIs1r2 ⊗ EEGs1r2	.79
fMRIs2r1 ⊗ EEGs2r1	.84
fMRIs2r2 ⊗ EEGs2r2	.84
Reliability: Incongruent Condition	
fMRI Context	
fMRIs1r1 ⊗ fMRIs1r2	.96
fMRIs2r1 ⊗ fMRIs2r2	.96
fMRIs1r1 ⊗ fMRIs2r1	.89
fMRIs1r2 ⊗ fMRIs2r2	.87
EEG Context	
EEGs1r1 ⊗ EEGs1r2	.98
EEGs2r1 ⊗ EEGs2r2	.98
EEGs1r1 ⊗ EEGs2r1	.91
EEGs1r2 ⊗ EEGs2r2	.87
Intercontextual: EEG/fMRI	
fMRIs1r1 ⊗ EEGs1r1	.82
fMRIs1r2 ⊗ EEGs1r2	.78
fMRIs2r1 ⊗ EEGs2r1	.86
fMRIs2r2 ⊗ EEGs2r2	.84
Reliability: Stroop Effect	
fMRI Context	
fMRIs1r1 ⊗ fMRIs1r2	n.s.
fMRIs2r1 ⊗ fMRIs2r2	n.s.
fMRIs1r1 ⊗ fMRIs2r1	n.s.
fMRIs1r2 ⊗ fMRIs2r2	n.s.
EEG Context	
EEGs1r1 ⊗ EEGs1r2	.80
EEGs2r1 ⊗ EEGs2r2	n.s.
EEGs1r1 ⊗ EEGs2r1	.70
EEGs1r2 ⊗ EEGs2r2	.64
Intercontextual: EEG/fMRI	
fMRIs1r1 ⊗ EEGs1r1	.78
fMRIs1r2 ⊗ EEGs1r2	.61
fMRIs2r1 ⊗ EEGs2r1	n.s.
fMRIs2r2 ⊗ EEGs2r2	n.s.

"n.s." means that the underlying correlation did not reach statistical significance

versions have shown that incongruent (font color and word meaning are different), as opposed to congruent (font color and word meaning are identical), task conditions produce longer response times when participants are asked to name the color of the written word (e.g., MacLeod, 1998; MacLeod & MacDonald, 2000). Numerous empirical confirmations of this effect, called the Stroop effect, have elevated this experimental design to the rank of a *psychological paradigm*. However, many questions remain unanswered with respect to the underlying cognitive and physiological bases of this prominent effect. To support either one of the contemporary theories about the basic mechanisms of Stroop performance (cf. Melara & Algom, 2003; Roelofs, 2003, 2010), the reproducibility of these data appears to provide a sufficient prerequisite.

As expected for the present data, at the group statistical level, differences between incongruent and congruent task response times were found quite consistently for almost all repeated measurement levels. Furthermore, Stroop-effect data could be shown to be consistently different from zero across almost all statistical tests. However, response times in the congruent and incongruent Stroop-task trials generally suggested better performance for the second than for the first measurement session, as well as for EEG as compared to MRI conditions. The data of particular interest, the Stroop-effect data, suggested a performance enhancement from the first to the second measurement session, which was even more present under MRI than under EEG conditions.

Stroop-task processing in different laboratory contexts across two measurement sessions

In general, response times could be shown to be longer in the first measurement session than in the second session three months later. Furthermore, response times were longer under MRI than under EEG laboratory conditions. In both laboratory contexts, the Stroop effect was significantly present, but it was reduced in the second measurement session (cf. Ackerman & Schneider, 1984; Beglinger et al., 2005; Davidson et al., 2003; Ellis & Dulaney, 1991). However, the performance improvement from session to session was significantly larger under MRI laboratory conditions. This implies that getting used to MRI measurement conditions is more difficult for participants than is getting used to EEG measurement conditions.

This might be explained by the different contextual properties in the respective laboratory environments, such as body position during measurement (lying in the MRI scanner vs. sitting upright in the EEG laboratory), preparation time before measurement (in the EEG laboratory, numerous electrodes must be placed, impedances must be adapted, and the electrode positions have to determined one after the other), and noise level during the measurement (gradient noise during fMRI vs. silence during EEG data measurement). It is, however, somewhat puzzling that there would be split-half intrasession inconsistencies in the first measurement session in the EEG context and in the second measurement session in the MRI context, even if the statistical power of the differences could be assumed to be moderate. Nevertheless, these results point to potential intertrial fluctuations in complex parameters such as the individual mean differences of incongruent and congruent response time values (cf. Overall & Woodward, 1975; Strauss et al., 2005).

In short, it appears that different laboratory contexts produce differential behavioral, and therefore physiological, effects. Hence, it is suggested that both the behavioral and physiological data obtained in different laboratory contexts should be interpreted in a complementary rather than in a redundant way. Effects of context have to be considered carefully. More detailed temporal analyses, as reported below, provide possible explanations for the observed inconsistency between laboratory contexts. Inconsistencies between study outcomes in general, potentially produced by different context variables, can also help to explain the reported differences across studies and model assumptions about the underlying mental and physiological processes causing Stroop interference effects (MacLeod, 2005; MacLeod & MacDonald, 2000).

Time-segment effects under different laboratory conditions: Intrasession variability in Stroop-task performance

Our results showed that for both measurement sessions in the EEG and MRI laboratory contexts, the Stroop effect was present for almost all time segments. Only in the second measurement session in the third time segment under MRI conditions did the Stroop effect disappear. This can be explained by the finding of a slight decrease in performance in the congruent and a slight improvement of performance in the incongruent task condition. Thus, the performance parameters of both task conditions converged. The behavioral data therefore suggest differential task-difficulty-dependent dynamics in performance levels within the different experimental sessions. However, it appears that this trend held not only for difficult task processing; it was also shown to be negatively influenced by noisy MRI-scanner conditions (Tomasi et al., 2005). In sum, the fluctuations of performance in both the incongruent and congruent task conditions during the first measurement session point to practice effects on task-unspecific motor responses or vigilance, but not to improved Stroop performance. However, Stroop performance did improve between sessions, potentially indicating learning effects based on long-term potentiation.

Intra- and intersession, and intermethodological, context-related reliability in Stroop-task performance

Following the notions of Golden and colleagues (1984), "true" reliability might only be assumed when both correlative (e.g., reliability) and quantitative (e.g., ANOVA and respective paired *t* tests for dependent variables) analyses confirm invariability between repeated measurement levels. Reliability measurements, however, are based on correlative logics, and therefore refer to consistency in the rank orders among individuals across repeated measurement levels (e.g., laboratory context, intrasession, and intersession repeated measurements), and not to quantitative differences. Theoretically, high consistency in rank orders between individuals can partner with significant differences in performance across measurement sessions, which was empirically confirmed by the present data. It is therefore questionable whether correlative parameters, even if they are generally used as a confirmation of retest consistency, provide an appropriate statistical approach.

Nevertheless, reliability has been set forth as one of the key features of scientific and clinical examination parameters to validly determine longitudinal effects. Indeed, Strauss and colleagues (2005) reported high test–retest reliability coefficients for the standard and emotional Stroop tasks when reliability was estimated separately for congruent and incongruent task-related response times. However, our Stroop-effect data did not show significant reliability for all repeated measurement levels. Thus, the reliability estimations presented here seem to partially confirm the results reported by Strauss et al. (2005; see also Eide, Kemp, Silberstein, Nathan, & Stough, 2002; Kindt, Biermann, & Brosschott, 1996; Siegrist, 1995, 1997); under MRI conditions, there was no significant reliability, and under EEG conditions, reliability was significant across only three of the four repeated measurement levels. Between laboratory contexts, only the reliability for the first measurement session was significant.

An explanation for these findings was discussed by Strauss et al. (2005) and originally presented by Overall and Woodward (1975). All of these researchers assumed that if two combined parameters are included to produce a new parameter—such as the Stroop effect, as a response time difference value between incongruent and congruent task performance—there is a possibility of compounding measurement errors from both underlying parameters. However, the question of whether between-individual or simply test-statistical reasons have produced the frequently reported inconsistencies in Stroop-effect data must be systematically investigated by means of appropriate experimental designs in subsequent studies.

It is strongly suggested that additional information be obtained, such as a debriefing about potential fluctuations in individually applied mental strategies, vigilance, or other confounding variables such as performance level in potentially related executive domains such as working memory (e.g., Kane & Engle, 2003; Unsworth & Engle, 2007). Furthermore, on the basis of the inconsistent results of reliability and ANOVA statistics, we suggest that pair-wise repeated measurement statistics be used, on the basis of ANOVA logics, to identify quantitative consistency with respect to potential individual fluctuations. These statistics should be used in the place of reliability estimations based on a correlative logic, which instead test for the consistency of rank orders of individuals across repeated measurement levels.

Final conclusions

We suggest that multiple structural changes for future studies on Stroop and comparable experimental designs may be useful, given the conclusions drawn from the present data:

1. Contextual variables should carefully be considered in both experimental single-shot and retest designs, but also when different studies on the same subject are compared.
2. Different laboratory contexts might have differential impacts on performance in both complex and simple tasks over time within one measurement session; it is therefore strongly recommended not to exert stress on study participants through overly long sessions, especially in demanding laboratory contexts such as the MRI scanner.
3. Contamination of the behavioral data by learning and/or practice effects must be carefully considered when psychological paradigms are used to confirm longitudinal changes in cognitive performance accompanied by respective neurodevelopmental plasticity. In the same manner, improvement of cognitive performance during neurorehabilitation must be controlled for general learning or long-term memory consolidation effects (i.e., long-term potentiation).
4. Choosing repeated measures ANOVA rather than correlative approaches such as reliability estimations is recommended to evaluate the reproducibility of effects.
5. For Stroop-task designs and similar paradigms, we suggest using congruent and incongruent response time measures separately, since differences in the individual means of both conditions may produce an overly complex construct with excessively complicated, comparatively unpredictable partial error variances (cf. Overall & Woodward, 1975; Strauss et al., 2005).
6. Experimental designs and clinical assessments should, if possible, be arranged in a such way that at least intrasession reproducibility of the core results can be assigned.
7. The results of single-shot studies should not be interpreted as a sufficient basis to verify or falsify theoretical frameworks without respective estimations confirming their reproducibility (cf. Button et al., 2013).

The present data support the view that Stroop interference processing, at least in the applied Stroop-task version here, is not a homogeneous phenomenon that is cognitively processed in a uniform or stable way. It appears that both group-related and individual fluctuations in performance parameters indicate corresponding fluctuations in the recruited mental subcomponents, such as working memory, memory, sustained attention, and/or monitoring, depending on fluctuations in the applied mental strategies (cf. Fehr, 2013). Therefore, different theoretical frameworks attempting to explain the underlying mental concepts behind Stroop interference processing (e.g., the "inadvertent-reading hypothesis," "horse-race theory," or "speed-of-processing" approaches; cf. MacLeod & MacDonald, 2000; Roelofs, 2010) may be equally suitable, depending on the individual sample characteristics, contextual factors, and the specifics of the applied Stroop design. It seems that further systematic experimental manipulations of similar Stroop designs will be necessary to isolate the underlying concepts of the Stroop-effect phenomenon.

Author note The authors are very thankful to the Center for Cognitive Sciences at Bremen University, Germany, for providing MRI scanner facilities, and to Manfred Herrmann for providing EEG facilities and the MRI scanner contingents of the Department of Neuropsychology at Bremen University and for several helpful suggestions. We also thank Susanne Hartfiel and Tessa Ayson for professional language editing.

References

Achtziger, A., Fehr, T., Oettingen, G., Gollwitzer, P., & Rockstroh, B. (2008). Strategies of intention formation are reflected in continuous MEG activity. *Social Neuroscience, 4,* 11–27.

Ackerman, P. L., & Schneider, W. (1984). Practice effects and a model for Stroop interference. Proceedings of the annual meeting of the Psychonomic Society, 1984, San Antonio, Texas, p. 289.

Beglinger, L. J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D. A., Crawford, J., & Siemers, E. R. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Archives* *of Clinical Neuropsychology, 20,* 517–529. doi:10.1016/j.acn.2004.12.003

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365–376.

Davidson, D. J., Zacks, R. T., & Williams, C. C. (2003). Stroop interference, practice, and aging. *Aging, Neuropsychology, and Cognition, 10,* 85–98.

Eide, P., Kemp, A., Silberstein, R. B., Nathan, P. J., & Stough, C. (2002). Test–retest reliability of the emotional Stroop task: Examining the paradox of measurement change. *Journal of Psychology, 136,* 514–520.

Ellis, N. R., & Dulaney, C. I. (1991). Further evidence for cognitive inertia in persons with mental retardation. *American Journal on Mental Retardation, 95,* 613–621.

Fahle, M., & Spang, K. (2003). Heterogeneity of brain responses to identical stimuli. *International Review of Sociology, 13,* 507–532.

Fehr, T. (2013). A hybrid model for the neural representation of complex mental processing in the human brain. *Cognitive Neurodynamics, 7,* 87–103. doi:10.1007/s11571-012-9220-2

Fehr, T., Achtziger, A., Hinrichs, H., & Herrmann, M. (2003). Interindividual differences in oscillatory brain activity in higher cognitive functions—Methodological approaches in analyzing continuous MEG data. In I. Reinvang, M. W. Greenlee, & M. Herrmann (Eds.), *The cognitive neuroscience of individual differences—New perspectives* (pp. 101–118). Oldenburg: bis Publishers.

Fehr, T., Wiedenmann, P., & Herrmann, M. (2006). Nicotine-Stroop and addiction memory—An ERP study. *International Journal of Psychophysiology, 62,* 224–232.

Golden, C. J., Sawicki, R. F., & Franzen, M. D. (1984). Test construction. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 19–37). New York, NY: Pergamon Press.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General, 132,* 47–70. doi:10.1037/0096-3445.132.1.47

Kindt, M., Biermann, D., & Brosschott, J. F. (1996). Stroop versus Stroop: Comparison of a card format and a single trial format of the standard color-word Stroop task and the emotional Stroop task. *Personality and Individual Differences, 21,* 653–661.

Lehrl, S. (2005). *Mehrfachwahl-Wortschatz-Intelligenztest (MWT-B)* (5th ed.). Balingen, Germany: Spitta.

MacLeod, C. M. (1998). Training on integrated versus separated Stroop tasks: The progression of interference and facilitation. *Memory & Cognition, 26,* 201–211. doi:10.3758/BF03201133

MacLeod, C. M. (2005). The Stroop task in cognitive research. In A. Wenzel & D. C. Rubin (Eds.), *Cognitive methods and their application to clinical research* (pp. 17–40). Washington, DC: American Psychological Association.

MacLeod, C. M., & MacDonald, P. A. (2000). Interdimensional interference in the Stroop effect: uncovering the cognitive and neural anatomy of attention. *Trends in Cognitive Sciences, 4,* 383–391.

Melara, R. D., & Algom, D. (2003). Driven by information: A tectonic theory of Stroop effects. *Psychological Review, 110,* 422–471. doi:10.1037/0033-295X.110.3.422

Overall, J., & Woodward, J. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin, 82,* 85–86.

Roelofs, A. (2003). Goal-referenced selection of verbal action: Modeling attentional control in the Stroop task. *Psychological Review, 110,* 88–125. doi:10.1037/0033-295X.110.1.88

Roelofs, A. (2010). Attention and facilitation: Converging information versus inadvertent reading in Stroop task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 411–422.

Siegrist, M. (1995). Reliability of the Stroop test with single-stimulus presentation. *Perceptual and Motor Skills, 81,* 1295–1298.

Siegrist, M. (1997). Test–retest reliability of different versions of the Stroop test. *Journal of Psychology, 13,* 299–306.

Strauss, G. P., Allen, D. N., Jorgensen, M. L., & Cramer, S. L. (2005). Test–retest reliability of standard and emotional Stroop tasks. *Assessment, 12,* 330–337.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643–662. doi:10.1037/0096-3445.121.1.15

Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., & Poline, J.-B. (2007). Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *NeuroImage, 35,* 105–120.

Tomasi, D., Caparelli, E. C., Chang, L., & Ernst, T. (2005). fMRI-acoustic noise alters brain activation during working memory tasks. *NeuroImage, 27,* 377–386.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114,* 104–132. doi:10.1037/0033-295X.114.1.104

Wöstmann, N. M., Aichert, D. S., Costa, A., Rubia, K., Möller, H.-J., & Ettinger, U. (2013). Reliability and plasticity of response inhibition and interference control. *Brain and Cognition, 81,* 82–94.