EXTRA
Dementia
and Geriatric
Cognitive Disorders

Dement Geriatr Cogn Disord Extra 2018;8:382–392

DOI: 10.1159/000493750
Received: July 24, 2018
Accepted: September 13, 2018
Published online: October 18, 2018

Original Research Article

Psychometric Properties of a Flanker Task in a Sample of Patients with Dementia: A Pilot Study

Lianne M.J. Sanders[a] Tibor Hortobágyi[a] Mala Balasingham[a]
Eddy A. Van der Zee[b] Marieke J.G. van Heuvelen[a]

[a]University of Groningen, University Medical Center Groningen, Center for Human Movement Sciences, Groningen, The Netherlands; [b]University of Groningen, Groningen Institute for Evolutionary Life Sciences (GELIFES), Groningen, The Netherlands

Keywords

Dementia · Inhibition · Assessment measures · Executive function · Flanker task · Stroop test

Abstract

Background/Aims: Reliable and valid neuropsychological tests for patients with dementia are scarce. To improve the assessment of attention and inhibitory control in dementia, we determined the feasibility, test-retest reliability, and validity of a Flanker task. ***Methods:*** Participants with all-cause diagnosed dementia (n = 22, mean age 84 years; mean Mini-Mental State Examination [MMSE] score = 19.4) performed a computerized Flanker task twice within 7 days. The Flanker task required participants to indicate the direction of target arrows flanked by congruent or incongruent arrows. Number of completed trials, accuracy, and reaction times (RTs) were recorded, and interference scores were calculated from basic scores. We examined the psychometric properties of the Flanker task and its relationship with the MMSE and Stroop test. ***Results:*** The Flanker task was feasible. Test-retest reliability was good for number of correct answers and RTs, and fair to poor for accuracy and the interference scores. The correlation of the Flanker task with Stroop and MMSE performance was fair to poor. ***Conclusion:*** The Flanker task appears to be feasible, and a reliable and valid measure of selective attention. Although the test-retest reliability for the Flanker RT interference measure was fair, future studies need to confirm its validity to measure inhibitory control in patients with dementia.

Lianne Maria Jantien Sanders
University of Groningen, University Medical Center Groningen
Center for Human Movement Sciences
Antonius Deusinglaan 1, NL–9713AV Groningen (The Netherlands)
E-Mail l.m.j.sanders@umcg.nl

Introduction

As the global population ages, the number of patients with dementia increases. Dementia is characterized by neuronal damage that leads to cognitive and motor impairments [1]. Up to 36% of patients with dementia show disinhibited behavior [2, 3] which becomes more prominent with disease progression [3]. Inhibitory control involves the intentional or unintentional suppression of unwanted actions in response to a stimulus [4]. Maintaining inhibitory control in dementia is a treatment target. There is uncertainty whether the inhibitory control tests that are valid and reliable in healthy adults adequately represent the level of inhibitory control of patients with dementia. Apart from global cognitive batteries that are designed specifically for patients with dementia, psychometric testing of neuropsychological measures is still uncommon in patients with dementia. Additionally, between-study comparisons of treatment effects on inhibitory control in dementia are limited by the large variation in the neuropsychological tests in use [5].

The Flanker task measures the ability to inhibit nonrelevant competing responses to a nonverbal stimulus [6]. While verbal communication becomes increasingly difficult for patients with dementia, the ability for nonverbal communication remains relatively preserved [7], making a Flanker task possibly suitable for this population. To the best of our knowledge, the psychometric properties of a Flanker task have not yet been established in patients with dementia. Here, we determined the feasibility, test-retest reliability, and validity of a computerized Flanker task. The results of the current study can be used to improve the assessment of inhibitory control in patients with dementia.

Materials and Methods

Subjects

Subjects were patients with dementia who participated in a multicenter study (Deltaplan Dementia, ZonMW: Memorabel 733050303) on the dose-response effects of exercise on cognition in patients with mild-to-moderate diagnosed dementia.

We obtained Flanker data in 22 participants 2 weeks after the end of the multicenter study (age = 83.8 ± 7.2 years; 11 women; median education = primary education + 2 years of lower secondary education). Participants were diagnosed with dementia by a primary care physician or geriatrician before inclusion in the exercise trial (*n* = 8 Alzheimer's disease [AD], *n* = 1 vascular dementia (VaD), *n* = 4 mixed AD+VaD, *n* = 9 unspecified). The Dutch College of General Practitioners advises to use the Diagnostic and Statistical Manual for Mental Disorders, ed 4 (DSM-IV) guidelines to diagnose dementia [8]. The average Mini-Mental State Examination (MMSE) score was 19.4 ± 5.0 with a range of 7–27. Participants were recruited from healthcare organizations that offered daycare or residential care facilities for patients with dementia in the Northern Netherlands.

Design

Each participant performed a computerized Flanker task twice with the re-test 7 days after first assessment. We compared the Flanker data with the MMSE [9] and Stroop task [10].

Measures

The Flanker task consisted of three conditions. In each condition, participants indicated the direction of the target arrow. In the congruent condition, a target arrow is flanked at each side by two nontarget arrows, which point to the same direction. In the incongruent condition, the target arrow is flanked by nontarget arrows, which point to the opposite direction. In the combined condition, congruent and incongruent trials were presented in a randomized order.

We programmed a computerized version with E-Prime 2.0 (Psychology Software Tools, Inc.). Participants sat in front of a computer monitor (at 70 cm distance) that was connected to a two-button box. Participants were asked to place their left and right index finger on the respective buttons and pressed the button corresponding to the direction of the target arrow. Stimuli were shown until the participant responded, and there was a fixed interval of 38 ms between the participant's response and a new stimulus. The number of performed trials in 45 s and the number of correct responses were recorded as well as the reaction times (RTs, s). We used this time limit of 45 s to restrict the burden of assessment for participants. The sequence of conditions was (1) congruent, (2) incongruent, and (3) combination. Participants completed five practice trials before each experimental condition. If a participant did not sufficiently understand the instructions after practice, another five practice trials were completed until the participant comprehended the instructions.

We compared the Flanker task with the MMSE and Stroop task. The MMSE is a global cognitive screening tool. Scores range from 0 to 30, and total score is used as outcome. The Stroop task measures selective attention and inhibitory control [10]. In the *word* condition (attentional processing), participants read the names of four colors (red, yellow, green, blue). In the *color* condition, participants named the colors. In the *color-word* condition (interference condition), participants were asked to name the color of words printed in incongruent colors. The number of total and correct responses within 45 s is recorded. We used this time limit of 45 s to restrict the burden of assessment for participants. The average RT for the Stroop task is acquired by dividing 45 s by the total number of responses. Accuracy (% correct) is calculated by dividing the number of correct responses by the total number of responses.

Level of inhibitory control was illustrated by the Stroop and Flanker interference scores. The method of calculating interference scores is given below.

Statistical Analyses

We used SPSS 22.0 (IBM Corp., Armonk, NY, USA) for data analyses with two-tailed significance set at $p < 0.05$. We calculated two interference scores for the Flanker task: (1) subtracting the mean RT for the correct congruent items from the mean RT for the correct incongruent items within the combination condition and (2) subtracting the mean accuracy for the congruent items from the mean accuracy for the incongruent items within the combination condition. Because the Stroop task was a paper-and-pencil test, we could not obtain RTs for correct and incorrect responses separately. Therefore, a Stroop RT interference score was calculated by subtracting the RT for total number of responses within the color condition from the RT for total number of responses in the color-word condition. A Stroop accuracy interference score was generated by subtracting the accuracy in the color condition from the accuracy in the color-word condition. For both measures, larger interference scores represent more interference.

The Flanker task was deemed feasible if all participants agreed to the assessment, were able to press the buttons and responded to the arrows, if there were no adverse events during testing and if accuracy scores were significantly better than chance. Accuracy scores were deemed better than chance if accuracy was ≥0.73. This cut-off score of ≥0.73 is based on a binomial distribution with the probability of correctly guessing being 0.5 and 22 trials (the average number of completed trials). In this situation, P(X ≥16 correct answers [as 16/22 = 0.73]) becomes <5%.

We determined the test-retest reliability of the Flanker task with paired *t* tests for differences in mean performance between the assessments, two-way random consistency single measures intraclass correlations (ICCs) with their confidence intervals (CIs) and Bland-Altman plots with exact CIs around the limits of agreement [11]. ICCs ≥0.9 were considered excellent, 0.75–0.9 good, 0.4–0.75 fair, and ≤0.4 poor test-retest reliability [12].

We determined the validity of the Flanker task by correlating the Flanker scores with MMSE and Stroop scores, and comparing the Flanker accuracy and RTs between the congruent and incongruent conditions.

In addition to a whole-group analysis, we stratified participants based on MMSE score. For this, we grouped participants around the average MMSE in lower-than-average ("low-MMSE") versus higher-than-average ("high-MMSE") subgroups.

Results

Scores on the MMSE and Stroop Reference Variables

The mean MMSE score was 19.4 ± 5.04. The mean number of total responses in the Stroop word, color, and color-word conditions was, respectively, 53.6 ± 21.75, 43.5 ± 18.71, and 21.0 ± 10.26. The mean number of correct responses in the Stroop word, color, and color-word conditions was, respectively, 53.5 ± 21.82, 42.7 ± 19.39, and 14.8 ± 10.98. The mean accuracy was 1.00 ± 0.01 for the word condition, 0.97 ± 0.05 for the color condition, and 0.69 ± 0.31 for the color-word condition. The mean Stroop RT interference score was 1.95 ± 2.54, and mean Stroop accuracy interference score was –0.29 ± 0.30.

Feasibility Flanker Task

All participants agreed to participate, were able to use the button box, and responded to the arrows. There were no adverse events. The mean number of practice trials was 24.8 on the first assessment and 20.7 at re-test. Stratified analyses based on MMSE score showed that participants with low MMSE needed five practice trials more than participants with high MMSE at first assessment (respectively, 27.5 vs. 22.5 trials, nonsignificant difference), but not at re-test (respectively, 20.5 vs. 20.8 trials).

Mean accuracy ranged from 0.76 to 0.92 (Table 1). The number of participants with accuracy scores ≥0.73 varied from 60 to 90% (Table 1). Stratified analyses based on MMSE score showed that specifically on the incongruent condition, accuracy was 22% lower for participants with low MMSE versus high MMSE.

The correlations between the number of completed trials and accuracy were fair to good ($r = 0.438$ to $r = 0.804$) in all conditions except the congruent condition at re-test ($r = 0.341$). The correlations between accuracy and RTs were all negative ($r = –0.203$ to $r = –0.682$), so there were no speed-accuracy trade-offs.

Test-Retest Reliability Flanker Task

On average, participants completed more trials in the congruent condition at retest compared with first assessment (Table 1), but test-retest performance was not significantly different for other conditions. ICCs were indicative of fair to good test-retest reliability for number of completed trials in all conditions, accuracy in the congruent condition only, and RT in all conditions (Table 1). ICCs indicated poor test-retest reliability for accuracy on the incongruent and combination condition. With respect to accuracy in the incongruent condition, stratified analyses based on MMSE score showed that accuracy scores were 0.21 (95% CI [–0.37, –0.06]) higher at re-test for participants with low MMSE but remained equal for participants with high MMSE. Stratified analyses based on MMSE score showed that ICCs for accuracy measures within the combination condition were on average –0.37 lower for participants with low MMSE versus high MMSE. The Bland-Altman plots for number of completed trials and accuracy (Fig. 1) show that differences in number of completed trials or accuracy between test and re-test did not depend upon the participants' mean performance.

DOI: 10.1159/000493750

© 2018 The Author(s). Published by S. Karger AG, Basel
www.karger.com/dee

Table 1. Test-retest reliability outcomes for Flanker performance

	First assessment mean (SD)	Second assessment mean (SD)	Second – first assessment mean (SD)	95% CI of difference second – first ass.	ICC (95%CI)
Flanker con.					
Trials completed, *n*	21.7 (10.5)	24.9 (10.6)	3.14 (5.67)	0.62, 5.65*	0.86 (0.68, 0.94)
Accuracy score, % correct	0.87 (0.14)	0.92 (0.10)	0.05 (0.12)	–0.00, 0.11	0.50 (0.11, 0.76)
Participants with accuracy ≥0.73, *n*	19/22	20/22	1	n.a.	n.a.
RT, s	2.30 (1.64)	1.83 (1.50)	–0.46 (1.06)	–0.93, 0.01	0.77 (0.53, 0.90)
Flanker incon.					
Trials completed, *n*	20.5 (11.2)	21.7 (9.72)	1.23 (4.24)	–0.65, 3.11	0.92 (0.81, 0.97)
Accuracy score, % correct	0.76 (0.24)	0.85 (0.18)	0.09 (0.24)	–0.02, 0.20	0.38 (–0.04, 0.69)
Participants with accuracy ≥0.73, *n*	13/22	17/22	4	n.a.	n.a.
RT, s	2.86 (2.29)	2.23 (1.72)	–0.64 (1.58)	–1.34, 0.07	0.71 (0.42, 0.87)
Flanker combi.					
Trials completed, *n*	19.8 (10.0)	21.3 (10.2)	1.55 (4.44)	–0.42, 3.51	0.90 (0.78, 0.96)
Accuracy score, % correct	0.76 (0.26)	0.84 (0.17)	0.08 (0.25)	–0.03, 0.20	0.37 (–0.51, 0.68)
Accuracy score con. items	0.77 (0.33)	0.88 (0.18)	0.10 (0.32)	–0.04, 0.24	0.25 (–0.19, 0.60)
Accuracy score incon. items	0.75 (0.28)	0.80 (0.23)	0.05 (0.28)	–0.07, 0.17	0.41 (–0.00, 0.70)
Participants with accuracy ≥0.73, *n*	14/22	18/22	4	n.a.	n.a.
RT, s	2.94 (2.52)	2.39 (1.85)	–0.54 (1.83)	–1.35, 0.27	0.66 (0.34, 0.84)
RT con. items	2.77 (2.45)	2.27 (1.88)	–0.50 (2.15)	–1.45, 0.45	0.52 (0.13, 0.77)
RT incon. items	3.20 (2.97)	2.62 (2.12)	–0.58 (2.16)	–1.54, 0.37	0.65 (0.32, 0.84)
RT correct con. items	3.01 (2.59)	2.29 (1.91)	–0.73 (2.26)	–1.73, 0.27	0.51 (0.12, 0.76)
RT correct incon. items	3.45 (3.54)	2.70 (2.54)	–0.75 (1.87)	–1.57, 0.08	0.82 (0.61, 0.92)
Flanker interference score RT					
Mean RT correct incon. items – mean RT correct con. items within combi. condition, s	0.43 (2.40)	0.42 (1.61)	–0.02 (2.23)	–0.96, 0.93	0.46 (0.05, 0.73)
Flanker interference score accuracy					
Mean accuracy incon. items – mean accuracy con. items within combi. condition, % correct	–0.03 (0.30)	–0.08 (0.22)	–0.05 (0.30)	–0.19, 0.08	0.36 (–0.07, 0.67)

con., congruent; incon., incongruent; combi., combination; RT, reaction time; ICC, intraclass correlation; n.a. = not applicable. * $p < 0.05$.

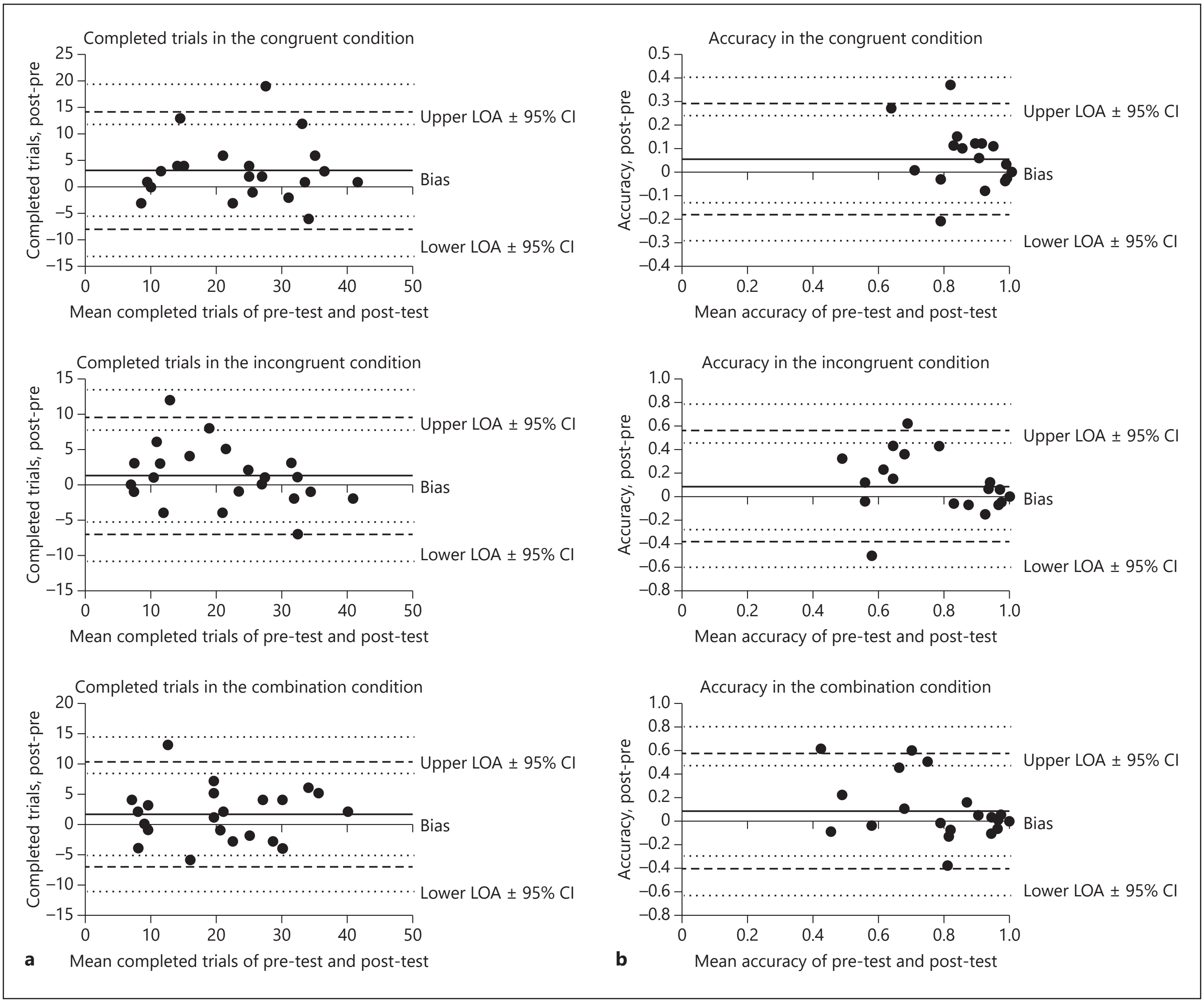

Fig. 1. a Bland-Altman plots for number of completed trials in all conditions. **b** Bland-Altman plots for accuracy in all conditions.

Within the combination condition, there were no test-retest differences in accuracy and RTs of participants on the congruent items and incongruent items separately (Table 1).

ICCs indicated fair test-retest reliability for the Flanker RT interference score, and poor test-retest reliability for the Flanker accuracy interference scores (Table 1).

Validity of the Flanker Task

Table 2 shows the output of the correlation analyses. There was a fair correlation of the Flanker number of completed trials on all conditions (Table 2) and accuracy on the congruent and incongruent condition (Table 2), with MMSE and Stroop word. Participants reacted on average 1.10 s slower (95% CI [0.29, 1.92]) in the incongruent compared with the congruent condition. There was a fair positive correlation between the Flanker and Stroop RT interference scores, and a fair negative correlation between the Flanker and Stroop accuracy interference scores (Table 2).

Table 2. Pearson correlations between Flanker, MMSE, and Stroop performance at first assessment

	MMSE	Stroop word number of completed trials	Stroop interference RT[a]	Stroop interference accuracy[b]
Flanker con. number of completed trials	0.489 (0.09, 0.76)	0.464 (0.05, 0.74)	–0.049 (–0.46, 0.38)	0.014 (–0.41, 0.43)
Flanker con. accuracy	0.536 (0.15, 0.78)	0.354 (–0.08, 0.68)	–0.318 (–0.65, 0.12)	–0.282 (–0.63, 0.16)
Flanker incon. number of completed trials	0.398 (–0.03, 0.70)	0.434 (0.02, 0.72)	–0.111 (–0.51, 0.33)	0.051 (–0.38, 0.46)
Flanker incon. accuracy	0.443 (0.03, 0.73)	0.274 (–0.17, 0.62)	–0.064 (–0.47, 0.37)	0.168 (–0.27, 0.55)
Flanker combi. number of completed trials	0.432 (0.01, 0.72)	0.441 (0.02, 0.73)	–0.089 (–0.49, 0.35)	–0.006 (–0.43, 0.42)
Flanker combi. accuracy	0.148 (–0.29, 0.54)	0.159 (–0.28, 0.54)	–0.201 (–0.57, 0.24)	–0.103 (–0.50, 0.33)
Flanker interference score RT[c]	–0.260 (–0.61, 0.18)	–0.600 (–0.82, –0.24)	0.502 (0.10, 0.76)	0.299 (–0.14, 0.64)
Flanker interference score accuracy[d]	–0.141 (–0.53, 0.30)	0.016 (–0.41, 0.44)	0.143 (–0.30, 0.63)	–0.680 (–0.89, –0.36)

con., congruent; incon., incongruent; combi., combination; RT, reaction time. [a] RT for total number of responses on color-word condition – RT for total number responses on color condition (s). [b] Accuracy color-word condition – accuracy color condition (% correct). [c] Mean RT correct incon. items – mean RT correct con. items within combi. condition (s). [d] Mean accuracy incon. items – mean accuracy con. items within combi. condition (% correct).

The accuracy and RT of participants on the Flanker task were equal for the congruent and incongruent items within the combination condition at first assessment (accuracy: difference congruent – incongruent items = 0.03, 95% CI [–0.10, 0.16]; RT: difference congruent – incongruent items = –0.43 s, 95% CI [–1.23, 0.37]).

Discussion

We investigated the psychometric properties of a 45 s Flanker task in a sample of patients with diagnosed dementia. We used this cut-off of 45 s per condition to restrict the assessment burden for our participants. In the current sample, this Flanker task was deemed feasible, and the test-retest reliability was good for number of correct answers and RT, and fair to poor for accuracy and the interference scores. The number of completed trials on the Flanker task correlated with MMSE and number of completed trials on the Stroop word condition. There were fair positive and negative correlations between the Flanker and Stroop RT and accuracy interference measures.

We deemed the Flanker task feasible if all subjects were able to perform the Flanker task without adverse reactions, which was confirmed. Also, accuracy scores had to be ≥0.73. The mean accuracy was ≥0.76 for all conditions, and 60–90% of participants obtained scores ≥0.73. Lower accuracy scores on the congruent and incongruent conditions were correlated with lower MMSE and stratified analyses based on MMSE score showed that specifically on the incongruent condition, accuracy was 22% lower for participants with low MMSE. Thus, the feasibility of the Flanker task may depend on the cognitive level of the participant. A limited understanding of the task, and difficulty memorizing the instructions, may contribute to suboptimal Flanker performance in patients with more severe dementia.

The test-retest reliability of the Flanker task was fair to good for number of completed trials and RT, but poor for accuracy on the incongruent and combination condition. Our results indicated that the Flanker task may be less reliable for patients with more cognitive impairments. Although a better performance at re-test could result from a number of factors, the finding that participants with low MMSE needed on average seven practice trials less at re-test as compared to first assessment (and contrary to participants with high MMSE) indicates the possibility of a learning effect in low-MMSE participants, or a quicker understanding of the instructions at re-test in low-MMSE participants. This warrants inclusion of a control group in future studies with the Flanker task. The poor test-retest reliability for accuracy on the incongruent and combination condition and the Flanker accuracy interference score makes the accuracy measure of the Flanker task less useful for clinical evaluation or evaluation of intervention effects. Test-retest reliability was fair for the Flanker RT interference score. This may render the Flanker RT interference measure preferable over the accuracy interference measure. However, considering the large CI, it is important to replicate these findings with a larger sample size. The use of Flanker RT interference measure is supported by previous research in a younger, healthy population [13].

We investigated the psychometric properties of the Flanker task using the Stroop test. In healthy adults, similar patterns of activation in the dorsolateral prefrontal cortex during a Stroop and Flanker task may indicate shared underlying interference processes [14]. The fair correlation between the Flanker and Stroop RT interference measures in our study may represent some of these shared underlying processes. However, there are noteworthy differences in neurocognitive processes during a Flanker versus Stroop task. In the Flanker task, the target stimuli and distractors are spatially apart, whereas in the Stroop task they are spatially integrated. Therefore, the perceptual interference may be stronger when performing a Stroop versus Flanker task, thereby increasing the error probability on incongruent trials on the Stroop task [14]. There may also be a higher demand for selective attention in the Stroop task, which was supported by a higher activation in inferior frontal regions compared with performance during a Flanker task in a meta-analysis of neuroimaging studies [14]. It remains undetermined whether the abovementioned findings also apply to a dementia population. Dementia-specific factors may further complicate the comparison of Flanker versus Stroop performance. Low test-retest reliability of the Flanker interference scores and difficulties with the Stroop task in a dementia population may have contributed to a low correlation between Flanker and Stroop performance. For example, color confusion [15] and impaired verbal fluency [7] may lower Stroop but not Flanker performance in patients with dementia. Also, we found a fair negative correlation ($r = -0.680$, Table 2) between the Flanker and Stroop accuracy interference measure. Visual inspection showed that there were 5 participants that performed better on the Flanker incongruent versus congruent items but not on the Stroop color-word versus color items. Higher within-person variability in performance on different cognitive tests is more common in patients with dementia compared with healthy peers [16]. Last, we included only correct responses in the Flanker RT interference measures, as opposed to inclusion of all responses in the Stroop RT interference measure. The abovementioned factors warrant a careful approach in the interpretation of our findings.

The current results show that participants were especially prone to slower RTs and more errors in the incongruent conditions of the Flanker task. These results support previous findings in patients with MCI and AD [17].

Because of the small sample size, the current study may have been underpowered. This must be particularly noted with respect to the stratified analyses based on MMSE scores, which need to be cautiously interpreted. In addition, it is a limitation of the current study that we were unable to use a gold standard to compare the Flanker task with, as no gold standard for measuring inhibitory control exists. Furthermore, we used a 45 s version of the Flanker task to restrict assessment burden for our participants, and caution is urged when generalizing the results to other Flanker tasks. Also, we used a pen-and-paper Stroop task which may not be completely comparable with a computerized task. We selected the MMSE and Stroop task because these measures are used regularly (MMSE) or fairly regularly (Stroop) in patients with dementia and cognitive impairment, despite their limitations [5]. For example, the Stroop test is used in exercise trials in these patient groups [18–21]. However, patients with dementia may have difficulty understanding and executing the task [22], which should be taken into account in the interpretation of the Stroop results. Furthermore, it should be noted that we assessed Flanker performance in subjects who participated in an exercise or control program for 6 months. However, it is unlikely that potential exercise effects confounded our results, because Flanker performance did not decrease from first to second assessment, and our unpublished data reveal no overall cognitive benefit of our exercise versus control program.

Conclusion

We need to build consensus on what measures to use in neuropsychological assessment for patients with dementia. As especially the Stroop task relies on different processes of selective attention, our results indicate that a Flanker task may be a feasible, reliable, and valid measure of attention in patients with dementia. The fair reliability of the Flanker RT interference score may indicate that the Flanker task may be suitable as a measure of inhibitory control in patients with dementia, although its validity needs to be confirmed in future studies. Not only a Flanker task, but other neuropsychological tests could be adapted to increase suitability for a dementia population. To optimize neuropsychological assessment for patients with dementia, it is important that researchers share successful and less successful attempts to create suitable dementia assessment tools.

Acknowledgements

The current study was supported by the Deltaplan Dementia (Deltaplan Dementia, ZonMW: Memorabel 733050303), the University of Groningen and University Medical Center Groningen. We thank Prof. Monicque Lorist from the University of Groningen for her contribution to the conception and design of the current method.

Statement of Ethics

The study was approved by the medical ethics committee of the University Medical Center Groningen (2014/523). Written informed consent was obtained for all participants. The study was conducted in accordance with the Declaration of Helsinki.

Disclosure Statement

The authors have no conflicts of interest to declare.

Author Contributions

Conception and design: L.M.J.S., M.B., M.J.G.v.H.; data acquisition: M.B.; analysis and interpretation: L.M.J.S., M.J.G.v.H.; drafting of manuscript: L.M.J.S., M.J.G.v.H., T.H.; revising the manuscript: all authors.

References

1 Raz L, Knoefel J, Bhaskar K. The neuropathology and cerebrovascular mechanisms of dementia. J Cereb Blood Flow Metab. 2016 Jan;36(1):172–86.
2 Mega MS, Cummings JL, Fiorello T, Gornbein J. The spectrum of behavioral changes in Alzheimer's disease. Neurology. 1996 Jan;46(1):130–5.
3 Lyketsos CG, Lopez O, Jones B, Fitzpatrick AL, Breitner J, DeKosky S. Prevalence of neuropsychiatric symptoms in dementia and mild cognitive impairment: results from the cardiovascular health study. JAMA. 2002 Sep;288(12):1475–83.
4 Faust ME, Balota DA. Inhibition, facilitation, and attentional control in dementia of the Alzheimer's type: The role of unifying principles in cognitive theory development. In: Gorfein DS, MacLeod CM, editors. Inhibition in cognition. Washington: American Psychological Association; 2007. pp. 213–38.
5 Bossers WJ, van der Woude LH, Boersma F, Scherder EJ, van Heuvelen MJ. Recommended measures for the assessment of cognitive and physical performance in older patients with dementia: a systematic review. Dement Geriatr Cogn Disord Extra. 2012 Jan;2(1):589–609.
6 Eriksen BA, Eriksen CW. Effects of noise letters upon the identification of a target letter in a nonsearch task. Percept Psychophys 1974;16(1):143–149.
7 Hubbard G, Cook A, Tester S, Downs M. Beyond words: Older people with dementia using and interpreting nonverbal behaviour. J Aging Stud 2002 May;16(2):155–167.
8 Moll van Charante E, Perry M, Vernooij-Dassen MJ, Boswijk DF, Stoffels J, Achthoven L, et al. NHG-Standaard Dementie (derde herziening). Huisarts Wet. 2012;55(7):306–17.
9 Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. 1975 Nov;12(3):189–98.

10 Stroop JR. Studies of interference in serial verbal reactions. J Exp Psychol Gen. 1992;121(1):15–23.
11 Carkeet A. Exact parametric confidence intervals for Bland-Altman limits of agreement. Optom Vis Sci. 2015 Mar;92(3):e71–80.
12 Andresen EM. Criteria for assessing the tools of disability outcomes research. Arch Phys Med Rehabil. 2000 Dec;81(12 Suppl 2):S15–20.
13 Clayson PE, Larson MJ. Psychometric properties of conflict monitoring and conflict adaptation indices: response time and conflict N2 event-related potentials. Psychophysiology. 2013 Dec;50(12):1209–19.
14 Nee DE, Wager TD, Jonides J. Interference resolution: insights from a meta-analysis of neuroimaging tasks. Cogn Affect Behav Neurosci. 2007 Mar;7(1):1–17.
15 Fisher LM, Freed DM, Corkin S. Stroop Color-Word Test performance in patients with Alzheimer's disease. J Clin Exp Neuropsychol. 1990 Oct;12(5):745–58.
16 Gamaldo AA, An Y, Allaire JC, Kitner-Triolo MH, Zonderman AB. Variability in performance: identifying early signs of future cognitive impairment. Neuropsychology. 2012 Jul;26(4):534–40.
17 Wang P, Zhang X, Liu Y, Liu S, Zhou B, Zhang Z, et al. Perceptual and response interference in Alzheimer's disease and mild cognitive impairment. Clin Neurophysiol. 2013 Dec;124(12):2389–96.
18 Castellano CA, Paquet N, Dionne IJ, Imbeault H, Langlois F, Croteau E, et al. A 3-Month Aerobic Training Program Improves Brain Energy Metabolism in Mild Alzheimer's Disease: Preliminary Results from a Neuroimaging Study. J Alzheimers Dis. 2017;56(4):1459–68.
19 Davis JC, Bryan S, Marra CA, Sharma D, Chan A, Beattie BL, et al. An economic evaluation of resistance training and aerobic training versus balance and toning exercises in older adults with mild cognitive impairment. PLoS One. 2013 May;8(5):e63031.
20 Liu-Ambrose T, Best JR, Davis JC, Eng JJ, Lee PE, Jacova C, et al. Aerobic exercise and vascular cognitive impairment: A randomized controlled trial. Neurology. 2016 Nov;87(20):2082–90.
21 Middleton LE, Black SE, Herrmann N, Oh PI, Regan K, Lanctot KL. Centre- versus home-based exercise among people with mci and mild dementia: study protocol for a randomized parallel-group trial. BMC Geriatr. 2018 Jan;18(1):27.
22 Burton RL, O'Connell ME, Morgan DG. Cognitive and Neuropsychiatric Correlates of Functional Impairment Across the Continuum of No Cognitive Impairment to Dementia. Arch Clin Neuropsychol. 2017 Nov 28:1–13. doi: 10.1093/arclin/acx112.