OXFORD UNIVERSITY PRESS | Archives of Clinical Neuropsychology

Construct Validity of the Stroop Colour-Word Test: Influence of Speed of Visual Search, Verbal Fluency, Working Memory, Cognitive Flexibility, and Conflict Monitoring.

Journal:	*Archives of Clinical Neuropsychology*
Manuscript ID	ACNP-2020-035.R1
Manuscript Type:	Original Empirical Article
Keyword:	Assessment, Attention, Executive functions, Test construction

Abstract

Eighty-five years after the description of the Stroop interference effect there is still a lack of consensus regarding the cognitive constructs underlying scores from standardized versions of the test. The present work aimed to clarify the cognitive mechanisms underlying direct (word-reading, colour-naming, and colour-word) and derived scores (interference, difference, ratio, and relative scores) from Golden's standardized version of the Stroop test. After a comprehensive review of the literature five target cognitive processes were selected for analysis: speed of visual search, phonological verbal fluency, working memory, cognitive flexibility, and conflict monitoring. These constructs were operationalized by scoring five cognitive tasks (WAIS- IV Digit Symbol, Phonological Verbal Fluency-letter A, WAIS-IV Digit Span, TMT B-A, and reaction times to the incongruent condition of a computerized Stroop task, respectively). Eighty-three healthy individuals participated in the study. Correlation and regression analyses were used to clarify the contribution of the five selected scores on the prediction of Stroop test scores. Data analyses revealed that Stroop word-reading reflected speed of visual search. Stroop colour-naming reflected working memory, and speed of visual search. Stroop colour-word reflected working memory, conflict monitoring, and speed of visual search. While the interference score was predicted by both conflict monitoring and working memory, it was the ratio score (i.e., colour-word divided by colour-naming) the one predicted by conflict monitoring alone. The present results will help neuropsychologists to interpret altered patient scores in terms of a failure of the cognitive mechanisms detailed here, benefitting from the solid background of preceding experimental work.

Key words: Assessment, Attention, Executive Functions, Test construction.

Introduction

The Stroop Colour-Word Test is considered one of the gold standards of attentional measures, and is one of the most widely used instruments in clinical and experimental neuropsychological settings (Strauss, Sherman, & Spreen, 2006). Among the different standardized versions, the test proposed by Golden (1978) is one of the most extensive owing to its relatively large number of specific norms for individuals from different socio-demographic conditions, and cultures (e.g., Lubrini et al., 2014; Strauss et al., 2006). This version features a three-page test booklet. On the first page, the words 'red', 'green', and 'blue', are printed in black ink and repeated randomly in columns (henceforth Stroop word-reading or SWR). On the second page, the item 'XXXX' appears repeatedly in columns, printed in red, green, or blue ink (henceforth Stroop colour-naming or SCN). On the third page (referred to as the interference page, henceforth Stroop colour-word or SCW), the words 'red', 'green', and 'blue' are printed in red, green, or blue ink but the words and the colours in which they are printed never match. The subject must look at each page and move down the columns, reading words or naming the ink colours as quickly as possible. The test yields three direct scores, based on the number of items completed on each of the three stimulus sheets in 45 seconds. In addition, interference (SCW – [(SWR*SCN)/(SWR+SCN)]; Golden, & Freshwater, 2002), difference (SCN-SCW), ratio (SCW/SCN), and relative ([(SCN-SCW)/SCN]*100) derived scores can be calculated (Lansbergen, Kenemans, & van Engeland, 2007; Scarpina, & Tagini, 2017; Sisco, Slonena, Okun, Bowers, & Price, 2016). However, current handbooks of neuropsychological assessment still assert that scores measured in core Stroop test conditions, such as SCW, have only marginal/acceptable reliability, and should not be used as the basis of diagnostic decisions, without supplementation by other data (Strauss et al., 2006). Clarifying this point is a central concern when using the Stroop test in experimental and clinical settings. In particular, it is crucial that the clinician has robust validation evidence regarding which cognitive operations underlie the scores provided by standardized neuropsychological measures. While most prior studies agree that the Stroop test has a complex and multifactorial structure comprising several cognitive mechanisms, there is a lack of consensus about their exact nature, and about their relative contribution to Stroop test scores. Table 1 presents an overview of 15 studies that have provided information useful to clarify the processes underlying Golden's version of the Stroop test (Golden, 1978).

[INSERT TABLE 1 ABOUT HERE]

The psychological constructs related to SWR and SCN have been largely related to verbal fluency, and speed of processing constructs. Beyond basic language skills such as verbal fluency (Lanham, Vanderploeg, & Curtiss, 1999) or reading skills (Protopapas, Archonti, & Skaloumbakas, 2007), the vertical arrangement of SWR and SCN stimuli in Golden´s version (100 words and 100 coloured 'XXXX' organized in five columns each with 20 words) appears to involve visual scanning abilities. In this regard, the association of both SWR and SCN, and neuropsychological tests with analogous visual scanning demands, such as TMT-A, TMT-B, Digit Symbol (WISC-III), or Block Design (WISC-III), supports the interpretation of SWR and SCN as useful measures of 'speed of visual processing' (Adrover-Roig, Sesé, Barceló, & Palmer, 2012; Bondi et al., 2002; Ríos, Periáñez, & Muñoz-Céspedes, 2004). However, it must be noted that only four of the 15 reviewed studies focused on the SWR score, with SCN and SCW attracting the most interest.

The psychological constructs related to the SCW score included cognitive control (Spikman, Kiers, Deelman, & van Zomeren, 2001), phonemic and semantic verbal fluency (COWAT; Bondi et al., 2002; Chaytor, Schmitter-Edgecombe, & Burr, 2006; Lanham et al., 1999; Llinàs-Reglà et al., 2015), processing speed (Symbol Digit, TMT-A, TMT-B, Finger tapping; Chaytor et al., 2006; Llinàs-Reglà et al., 2015; Ríos et al., 2004; Sánchez-Cubillo et al., 2009; Spikman et al., 2001), interference control (Ríos et al., 2004), behavioural disinhibition (NPI-Disinhibition, Heflin et al., 2011), cognitive flexibility (WCST-Perseverative errors, TMT-B, B-A, Chaytor et al., 2006; Sánchez-Cubillo et al., 2009), or working memory (Digit Forwards, Digit Backwards; Adrover-Roig et al., 2012; Llinàs-Reglà et al., 2015; Protopapas et al., 2007; Sánchez-Cubillo et al., 2009). However, some of the aforementioned associations are open to question, given that many of these neuropsychological measures require more than a single cognitive ability. For example, the association between SCW and factors like cognitive flexibility, based on its correlation with the TMT-B score (Chaytor et al., 2006), disappeared after controlling for visual search and perceptual speed factors in a multiple regression analysis, which raises questions regarding the involvement of this mechanism in SCW performance (Sánchez-Cubillo et al., 2009). Also, the association between SCW and verbal fluency, as measured by FAS, may be more related to shared executive control abilities such as working memory or shifting/updating than to linguistic skills, as revealed by correlations between FAS and other tasks involving working memory (Aita et al., 2018).

The psychological constructs related to derived test scores are much less clear. For instance, two studies associated the Golden's interference score to response inhibition (commission errors in TOVA, WCST total errors; Cox et al., 1997; Kluttz, & Golden, 2016), and one more with executive functions (Sisco et al., 2016). Also, the difference score has been related to sustained selective processing (Serial Subtraction task), visuo-motor scanning (Digit Symbol, Letter Cancelation, SDMT, TMT-A, TMT-B; Shum et al., 1990), and reading skills (Protopapas et al., 2007). Only the study by Sisco, et al., (2016) reported data about a relationship of the ratio and relative scores with other executive function measures. Particularly, while the ratio score correlated with both TMT ratio and WCST completed categories, the relative score correlated with the TMT ratio. However, bearing in mind the relatively scant available data, it is difficult to extract any robust conclusion for validity purposes.

In addition to the analysis of prior neuropsychological evidence, the study of the association between the Stroop pencil-and-paper test and computerized versions may help to clarify the underlying cognitive processes for at least two sound reasons. First, computerized versions reduce many of the cognitive demands associated with the standard test. For instance, Stroop stimuli are presented one by one in the centre of a screen, minimizing both visual search demands, and the potential interference from flanking words when targets are arranged in vertical rows. Also, responses in computerized tasks use to be made by pressing a button reducing certain verbal demands. In spite of this, the interference effect is still present in computerized versions (MacLeod, & MacDonald, 2000). Secondly, computerized Stroop tasks accumulate a large amount of behavioural, physiological and neuroimaging data that supports an association between the Stroop reaction time (RT) interference effect and a more specific executive control mechanism, i.e., conflict monitoring, or the operation of a system that detects the occurrence of conflicts in information processing and then passes information on to centres responsible for control (e.g., Botvinick, Braver, Barch, Carter, & Cohen, 2001; Botvinick, Cohen, & Carter, 2004). However, and as mentioned above, only few studies have analysed the association between computerized Stroop tasks and neuropsychological data from pencil-and-paper versions, providing contradictory results (Kindt, Bierman, & Brosschot, 1996; Penner et al., 2012). This is a relevant issue given the theoretical and applied interest of building a bridge between these two sets of evidences.

The present study aimed to improve the existing knowledge about construct validity of Stroop Test scores (Golden, 1978). Specifically, there are discrepancies regarding the involvement of some of these cognitive abilities in Stroop scores. There is also a lack of consensus regarding the terminology used to refer to cognitive constructs. In addition, there has been little effort to connect theoretical advances from the experimental literature with existing neuropsychological data about the standardized version of the Stroop task. At least, some of these problems can be attributed to the extended use of bivariate correlational methodologies in validation studies, that makes it difficult to identify the relative contribution of different cognitive processes to the relationship observed between two scores. Alternatively, the use of multiple regression provides specific data regarding the joint and unique contribution of the different predictors to the variance of a given criteria score. In an attempt to overcome these limitations, in the present work, five well-known cognitive variables previously associated with Stroop test performance (i.e., processing speed, verbal fluency, working memory, cognitive flexibility, and conflict monitoring) were introduced in a multiple regression model as predictors of direct and derived scores from the Golden's version of the Stroop test.

Methods

Participants

A sample of 83 healthy adults (mean ± SE age = 25.2 ± 1.1; range= 18-64; mean ± SE years of education =13.6 ± .3; range= 8-24; 61 females) took part in this study. Participants were recruited as volunteers from university courses, university staff, and health care centres. Each participant provided a self-reported history of medical and psychiatric problems. Exclusion criteria were a history of neurological disease, psychiatric illness, head injury, stroke, substance abuse (excluding nicotine), learning disabilities, or any other difficulty that could interfere with testing. All participants had normal or corrected-to-normal vision.

Instruments and Procedure

A neuropsychological examination was conducted by experienced psychologists in one session that included a brief interview, standardized neuropsychological testing, and computerized testing. This study was completed in compliance with institutional research standards for human research and in accordance with the Declaration of Helsinki.

Stroop Test: The Spanish adaptation of the Stroop test (Golden, 1994) was used. The number of correct responses in 45 seconds in the word-reading (SWR), colour-naming (SCN), and colour-word (SCW) conditions was recorded. The examiner indicated the errors, and participants were asked to correct them before continuing. Four derived scores were also calculated. First, the interference score proposed by Golden and Freshwater (2002) represents the difference between SCW score and SCW′, where SCW′ equals (SWR*SCN)/(SWR+SCN). Second, a difference score represents the difference between SCN and SCW conditions. A ratio score was calculated by dividing the SCW score by the SCN score. Lastly, a relative score was calculated according to the formula [(SCN-SCW)/SCN]*100 (see a review of Stroop interference derivate scores in Lansbergen et al., 2007, and Sisco et al., 2016). Evidences of compatibility between English monolinguals and Spanish monolinguals performance of Stroop Test have been shown in preceding cross-linguistic studies, revealing a lack of differences in the scores between samples (Rosselli, Ardila, Santisi, et al., 2002).

Digit Symbol Subtest (WAIS-IV): Speed of visual search was assessed using the Digit Symbol subtest from the Spanish adaptation of the WAIS-IV (Wechsler, 2012). This score has shown the highest load in the Processing Speed factor as described in the WAIS-IV construct validity data (Wechsler, 2012). The score in the test (number of symbols correctly encoded in 2 minutes) was considered the variable for analyses.

Digit Span (WAIS-IV): Digit Span subtest from the Spanish adaptation of the WAIS-IV (Wechsler, 2012) was used to assess working memory. This test was selected because it shown the highest load in the Working Memory factor as described in the WAIS-IV construct validity data (Wechsler, 2012). Validation studies have shown that it loaded together with other verbal working memory tasks on working memory latent variables (Kane et al., 2004). The Digit Span score (the sum of the number of correctly recalled items in both Digit Forward and Digit Backward tasks) was recorded as the variable for analysis.

Trail Making Test B-A: Cognitive flexibility was assessed using the TMT B-A difference score. Parts A and B of the Trail Making Test, were administered to participants according to the guidelines presented by Strauss et al. (2006). TMT B−A score (time difference in seconds between part B and part A) was the variable considered for analysis. Prior TMT validation studies have suggested that TMT B-A represent a relatively pure measure of cognitive flexibility/task-switching abilities on the basis of its association with the behavioural switch cost as measured in a computerized WCST-like paradigm (Sánchez-Cubillo et al., 2009).

Phonemic Verbal Fluency (letter A): According to the guidelines presented by Strauss et al. (2006), participants were asked to produce as many words beginning with the letter 'A' as quickly as possible during a period of one minute, excluding proper names or the same word with different ending. The number of admissible different words produced in the selected phonological category was the score considered as the variable for analysis (or Fluency A), given that the letter A loads more on verbal fluency latent variables besides alternative ones (Whiteside et al., 2016). Evidences of compatibility between English monolinguals and Spanish monolinguals performance of Phonemic Verbal Fluency-letter A score have been shown in cross-linguistic studies, revealing a lack of differences between samples (Rosselli, Ardila, Salvatierra, et al., 2002).

Computerized Stroop task: This task, inspired by experimental paradigms used in previous behavioural and brain activation studies (Mead et al., 2002; Penner et al., 2012; Swick, & Jovanovic, 2002), was used in order to measure conflict monitoring (Botvinick et al., 2001; Botvinick et al., 2004). Participants were instructed to name to the ink colour of the word appearing in the centre of a computer screen as quickly and accurately as possible. Given that it has been demonstrated that the Stroop interference effect has been affected both by the proportion of congruent an incongruent trials (being smaller when incongruent trials are frequent), and the duration of the inter-stimulus-interval (disappearing during very short intervals; De Jong, Berendsen, & Cools, 1999), both parameters were manipulated to better capture the interference effect. Accordingly, the task consisted of 144 trials from one of three experimental conditions randomly intermixed (48 trials each): congruent colour words, incongruent colour words, and colour-neutral words. In the congruent condition three colour words (red, green, and blue) appear in a congruent ink colour. In the incongruent condition the same three colour words appear in one of three incongruent ink colours (red, green, or blue). Lastly, in the colour-neutral condition, two Spanish non-colour words (i.e., 'glasses', 'table', in Spanish) appear in one of three ink colours (red, green, and blue). Responses were made by pressing one of three buttons on a computer keyboard with the index finger of the dominant hand in an array corresponding to the three possible printed colours (red, green, and blue). The stimulus duration was 150 milliseconds. A long inter-stimulus-interval of 2400 milliseconds was established. A practice block of 18 trials (6 in each condition) was administered to the participants before the task. Given the already described problems when selecting control conditions to estimate the interference effect by means of subtraction methods (either when using neutral or congruent conditions; Regan, 1978; MacLeod, 1991), reaction times to the incongruent condition (Incongruent RT) were used as the dependent measure for analysis. The use of this variable in a regression context provides a better measure of interference effects, since commonalities between the different predictors can be estimated and controlled (See Methods).

Data Analyses

A set of exploratory correlation analyses helped to describe the pattern of relationships within Stroop test variables, and those between Stroop test variables and the remaining cognitive scores. Multiple regression analyses (stepwise) were performed for each Stroop test score using the five selected cognitive scores as potential predictors (i.e., Digit Symbol, Digit Span, Fluency A, TMT B-A, and Incongruent RT). In order to estimate and control the potential effect of age and education in the modulation of the relationships between predictors and criteria scores, these two variables were also introduced in the regression models together with the five cognitive scores. Standard criteria were applied for a predictor variable to be included (probability of $F < .05$) or excluded (probability of $F > .1$) from each model. Normality, linearity of residuals, absence of multicollinearity, and independence of errors were assessed prior to all regression analyses. *A priori* contrasts were used in all statistical comparisons with an uncorrected significance level of $p < .05$ given that our variable selection derived from a review of studies that had already demonstrated a relationship between scores. The SPSS v.22.0 statistical software package was used to perform analyses. G*Power statistical software (Faul, Erdfelder, Buchner, & Lang, 2009) was used to estimate both the effect sizes (f^2) and the statistical power of the analyses.

Results

Descriptive statistics of all scores are shown in Table 2.

[INSERT TABLE 2 ABOUT HERE]

Exploratory Correlation Analyses

Intercorrelation Pearson coefficients between Stroop scores and other cognitive measures are shown in Table 3. All direct scores significantly correlated to each other. While all derived scores had significant correlations with each other, and with the SCW condition, only the difference score was also related to the SCN score (see Table 3). With regard to the remaining cognitive scores, SWR correlated with Digit Symbol. The SCN score correlated with Digit Symbol, Digit Span, and Incongruent RT. SCW correlated with Digit Symbol, Digit Span, Incongruent RT, and marginally ($p= .056$) with Fluency A score. The interference score correlated significantly with Digit Symbol, Digit Span, and Incongruent RT, while ratio and relative scores only correlated with Incongruent RT. The difference score showed no relation with any of the considered cognitive variables. The linear dependency measured between the ratio and relative scores led to exclude the latter in subsequent analyses.

[INSERT TABLES 3 AND 4 ABOUT HERE]

Regression Analyses

The assumptions of regression were meet for all the analyses being performed. Thus, visual inspection of histograms and partial regression plots revealed that normality and linearity of residuals were met in all cases. There was an absence of multicollinearity between the independent variables of all regression analyses (FIV < 1.13 in all cases). The analyses also revealed that errors of prediction were independent of each other in all regression models (Durbin-Watson values between 1.5 and 2.1 in all cases).

The multiple regression performed on SWR score as the criterion was significant (R^2 = .049; p < .045; Effect size f^2 = .05; Power = .25). Digit Symbol was the variable entering the model with a significant contribution of 4.9% to the prediction of SWR (see Table 4, top panel). Age and education showed non-significant contributions to the prediction of SWR with 3.2 and 1.4% of the variance, respectively (ps > .107).

The multiple regression performed on SCN score as the criterion was significant (R^2 = .132; p < .01; Effect size f^2 = .18; Power = .78). Digit Symbol, and Digit Span were the variables entering the model, accounting for 8%, and 8.5% of the variance of SCN, respectively (see Table 4, second panel). Age and education showed non-significant contributions to the prediction of SCN with 0.3 and 0.2% of the variance, respectively (ps > .635).

The multiple regression performed on SCW score as the criterion was significant (R^2 = .24; p < .027; Effect size f^2 = .32; Power = .97). Digit Symbol, Digit Span, and Incongruent RT were the variables entering the model, accounting for 6%, 9.9%, and 6% of the variance of SCW, respectively (see Table 4, third panel). Age and education showed non-significant contributions to the prediction of SCW with 0.05 and 0.003% of the variance, respectively (ps > .830).

The multiple regression performed on Stroop interference score score as the criterion was significant (R^2 = .123; p < .04; Effect size f^2 = .14; Power = .65). Digit Span, and Incongruent RT were the variables entering the model, accounting for 5.2%, and 7% of the variance of Stroop interference score, respectively (see Table 4, fourth panel). Age and education showed non-significant contributions to the prediction of Stroop interference with 2.3 and 0.2% of the variance, respectively (ps > .169).

Lastly, the multiple regression performed on Stroop ratio score as the criterion was significant ($R^2 = .06$; $p < .026$; Effect size $f^2 = .06$; Power = .30). Incongruent RT was the variable entering the model, accounting for 6% to the prediction of Stroop ratio score (see Table 4, bottom panel). Age and education showed non-significant contributions to the prediction of Stroop interference with 2 and 0.1% of the variance, respectively (ps > .205).

Discussion

The aim of this study was to clarify which cognitive mechanisms underlie Stroop standardized test scores (Golden, 1978). A sample of 83 healthy individuals was assessed by means of a battery of neuropsychological tests and a computerized task that, according to a comprehensive review of the literature, had previously demonstrated a relationship with Stroop test performance.

A series of exploratory Pearson product-moment correlations confirmed the relationship between Stroop direct scores, supporting the general assumption of common underlying cognitive mechanisms. As shown in Table 3, results also confirmed our *a priori* assumption about a relationship between Stroop scores and most cognitive measures selected for the analyses. A series of multiple regression analyses using Stroop direct (SWR, SCN, SCW) and derived scores (interference and ratio) as the criteria, were performed to assess the predictive value of the five selected cognitive scores measuring speed of visual search, phonological verbal fluency, working memory, cognitive flexibility, and conflict monitoring. In the following section the results will be discussed in relation to preceding findings.

Stroop Test Direct Scores

In summary, multiple regression analysis of SWR suggested that this score was predicted by speed of visual search, as measured by WAIS-IV Digit Symbol score (accounting for a 4.9% of the variance). Likewise, multiple regression analyses performed on SCN suggested that this score was primarily predicted by working memory, as measured by WAIS-IV Digit Span score, and secondarily by speed of visual search, as measured by WAIS-IV Digit Symbol score (accounting together for a 16,5% of the variance). Lastly, multiple regression analysis performed on SCW suggested that working memory, conflict monitoring, and speed of visual search were the main contributing variables (as measured by Digit Span, Incongruent RT, and Digit Symbol, respectively), accounting for 21,9% of the variance. The implications of these results and their association with preceding literature are discussed below.

The association between all the Stroop test direct scores and processing speed is coherent with preceding studies in healthy controls as well as clinical populations (e.g., Adrover-Roig et al., 2012; Bondi et al., 2002; Llinàs-Reglà et al., 2015; Ríos et al., 2004). Also, the association between working memory and both SCN and SCW test conditions has been replicated in a number of studies (Adrover-Roig et al., 2012; Bondi et al., 2002; Llinàs-Reglà et al., 2015; Sánchez-Cubillo et al., 2009). Even computational proposals of the Stroop task have modelled the way task demand specifications must provide an input to the system in order to perform the appropriate task (i.e., ‘respond to colour’ or ‘respond to word’; Cohen, Dunbar, & McClelland, 1990). In addition, the association between SCW and both working memory and conflict monitoring is compatible with the results described by Kane and Engle (2003). In their five experiments, they showed that individual differences in working memory predicted Stroop performance, and suggested that interference was jointly determined by working memory (i.e., goal maintenance), and executive control mechanisms (i.e., competition resolution). As detailed above in the introduction to this article, the interpretation of conflict monitoring as one key executive control mechanism underlying the Stroop interference effect, relies on a considerable amount of behavioural, electrophysiological, and neuroimaging preceding evidence. Consequently, the fact that computerized Stroop RTs in the interference condition (Incongruent RT) was a reliable predictor of the SCW score here, could be taken as an evidence of a plausible implication of conflict monitoring mechanisms in the score of the standardized test. Recently, attempts have been made to refine the type of conflict underlying Stroop performance. For instance, Kalanthroff, Goldfarb, & Henik (2013) showed that both informational conflict (the conflict between contradictory information that arises from the irrelevant word meaning and the relevant word colour) and task conflict (the conflict between two tasks -the relevant colour naming task and the irrelevant word reading task) appear in the Stroop task. However, only the informational conflict seems to determine the interference effect (Kalanthroff et al., 2013). In any case, the possibility of conflict monitoring being a significant mechanism underlying Stroop performance does not invalidate the possibility of complementary cognitive mechanisms being implicated (MacLeod & MacDonald, 2000).

The present analyses did not replicate the association between SCW and cognitive flexibility (as measured by TMT B-A; i.e., Chaytor et al., 2006; Llinàs-Reglà et al., 2015; Sánchez-Cubillo et al., 2009). In light of the present results, this association may owe more to common speed and working memory factors than to cognitive flexibility. This latter explanation is in line with previous data. For instance, factorial studies revealed that the SCW score loaded on factors that appear to represent processing speed (e.g., Digit Symbol, TMT A, FAS) rather than executive functions as measured by tests such as WCST and TMT B (Bondi et al., 2002; Boone, Ponton, Gorsuch, Gonzalez, & Miller, 1998). Also, Sánchez-Cubillo et al. (2009) showed a correlation between SCW and TMT B-A (r= -.31), which disappeared when processing speed and working memory were entered in their multiple regression analyses. Taken together, the present correlation and regression results suggest that the association between TMT B-A and SCW may rely more on common working memory processes, rather than cognitive flexibility, which make a secondarily contribute to both scores.

The results revealed a marginally significant correlation between phonological verbal fluency (i.e., Fluency A), and SCW condition, but not with SWR or SCN conditions. Moreover, when a multiple regression was performed on SCW, the influence of phonological verbal fluency disappeared in the presence of the remaining variables. Contradictory evidence exists in the literature about the relation between Stroop test scores and verbal fluency with both positive (Chaytor et al., 2006; Lanham et al., 1999; Llinàs-Reglà et al., 2015) and negative findings (Adrover-Roig et al., 2012; Sisco et al., 2016). The present results support the view that verbal fluency does not seem to be a key variable to account for Stroop test performance. Even if apparently surprising, fluency tasks require participants to select freely among a number of potential responses that would be minimized or even absent during Stroop test conditions (Botvinick et al., 2001).

Stroop Test Derived Scores

First, and regarding Golden's interference score, only six of the 15 reviewed studies considered this score for analysis. Importantly, only the investigations by Cox et al. (1997) and by Kluttz & Golden (2016) associated the interference score to a specific cognitive operation, i.e., response inhibition, on the basis of its relationship with commission errors in the TOVA, and WCST total errors, respectively. However, the lack of association between Stroop interference effects and different inhibition tasks like the go/no go task (Christ, Holt, White, & Green, 2007), the negative priming task (Vitkovitch, Bishop, Dancey, & Richards, 2002), or the Hayling test (Cipolotti et al., 2016) also suggests that inhibition may be an inadequate construct to account for the above mentioned effects. In fact, both stop-signal tasks (like the TOVA), and task-switching paradigms (like the WCST), have been related to various coordinated cognitive operations (e.g., Lange, Kröger, Steinke, Seer, Dengler, & Kopp, 2016; Verbruggen & Logan, 2008), making it difficult to clarify which of them may determine the association with Stroop test scores. The present data, showing an association between Stroop Interference and both conflict monitoring and working memory, as measured by more specific tasks, may provide a parsimonious account for both present and previous validation data given the implications of conflict monitoring and working memory in both the WCST (Periáñez, Maestú, Barceló, Fernández, Amo, & Ortiz Alonso, 2004; Lange, et al. 2016) and the stop-signal task (Verbruggen & Logan, 2008). Second, and regarding the ratio score, only one reviewed study (Sisco et al., 2016) considered it for analysis, reporting a significant association with 'executive function', as measured by both the TMT ratio, and WCST completed categories. The present data are compatible with such a general interpretation, and allows clarifying that it was conflict monitoring, and not working memory or cognitive flexibility (that did not entered the regression model as significant predictors), the key executive mechanism involved in Stroop ratio score. Lastly, it is important to notice that while the difference score was associated to sustained selective processing (Serial Subtraction task), visuo-motor scanning speed (Digit Symbol, Letter Cancelation, SDMT, TMT-A, TMT-B; Shum et al., 1990; Sisco et al., 2016), and reading skills (Protopapas et al., 2007) in three previous works, our analyses were unable to find any association with the cognitive scores being considered. It has to be noticed that the cognitive variables used here differed from those used in preceding investigations. However, the fact that the difference score was the only derived index showing no relationship with any non-Stroop cognitive score, may provide support to the idea of Lansbergen et al., (2007) that the difference score suffers from computational problems making it inadequate for clinical purposes.

In summary, the present data clearly suggests that the SWR requires speed of visual search, SCN primarily reflects working memory and secondarily speed, while SCW reflects working memory, conflict monitoring, and speed. The results also suggest that the analysed derived interference indexes (i.e., interference and ratio scores) minimize speed of visual search. In this regard, while Golden's interference score reflected both working memory and conflict monitoring, the ratio score resulted to be the purest marker of conflict monitoring. However, two potential limitations should be highlighted. First, it is important to notice that the amount of variance accounted by certain predictors, although statistically significant, was modest in some cases. The analysis of the possible reasons of these effects may help clarifying whether it represents a potential limit to the interpretative utility of the current data. One explanation could be that cognitive abilities others than those considered here might be also playing a role in Stroop test performance. This possibility, although plausible, was minimized in the present work by means of a detailed analysis of preceding literature. A complementary possibility is that the different cognitive demands modulating Stroop performance may be operating not as mere additive factors (i.e., the Donders' fallacy of "pure insertion"; Jensen, 2006), but as interacting factors that will increase execution time beyond the time taken by each individual operation alone. In fact, the multiple regression methodology can clearly estimate the association between a group of predictors and a given criteria, but it is unable to estimate the way the cognitive processes underlying those predictors will interact in the cognitive system when operating together (unless another variable measuring this interaction is introduced in the model). If this later possibility is true, the variables accounting for modest amounts of variance should not be neglected when interpreting Stroop scores. In fact, the apparently small effects could be reflecting "the tip of the iceberg" of a more robust association between predictors and outcomes, unable to be captured by current regression models. While contrasting this later hypothesis exceeds the goals and methods described here, future studies investigating the validity of a complex tool like the Stroop test, should consider this possibility. In any case, caution must be taken when using the present validation data in applied contexts, especially with those Stroop scores in which the portion of accounted variance was modest. Second, it should be noted that the number of female participants was higher than the number of males. This, however, should not be considered an important bias in the present work, since sex has demonstrated no significant influence modulating Stroop test scores in at least three different Spanish samples from a preceding normative study (Lubrini, et al., 2014). In spite of it, the present results on the Stroop test validity fill an important gap in the literature (Strauss et al., 2006), and will help researchers and clinicians to interpret altered patient scores in terms of a failure of the cognitive mechanisms detailed here, benefitting from the solid background provided by preceding experimental work on Stroop tasks.

ACKNOWLEDGMENTS:

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Adrover-Roig, D., Sesé, A., Barceló, F., & Palmer, A. (2012). A latent variable approach to executive control in healthy ageing. *Brain and Cognition, 78*, 284-299. https://doi.org/10.1016/j.bandc.2012.01.005

Aita, S. L., Beach, J. D., Taylor, S. E., Borgogna, N. C., Harrell, M. N., & Hill, B. D. (2018). Executive, language, or both? An examination of the construct validity of verbal fluency measures. *Applied neuropsychology. Adult, 7*, 1-11. https://doi.org/10.1080/23279095.2018.1439830

Bondi, M. W., Serody, A. B., Chan, A. S., Eberson-Schumate, S. C., Delis, D. C., Hansen, L. A., & Salmon, D. P. (2002). Cognitive and neuropathologic correlates of Stroop Color-Word Test performance in Alzheimer's disease. *Neuropsychology, 16,* 335–343. http://dx.doi.org/10.1037/0894-4105.16.3.335

Boone, K. B., Ponton, M. O., Gorsuch, R. L., Gonzalez, J. J., & Miller, B. L. (1998). Factor analysis of four measures of prefrontal lobe functioning. *Archives of Clinical Neuropsychology, 13,* 585–595. https://doi.org/10.1093/arclin/13.7.585

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*, 624-652. https://doi.org/10.1037/0033-295X.108.3.624

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences, 8*, 539-546. https://doi.org/10.1016/j.tics.2004.10.003

Chaytor, N., Schmitter-Edgecombe, M., & Burr, R. (2006). Improving the ecological validity of executive functioning assessment. *Archives of Clinical Neuropsychology, 21*, 217-227. https://doi.org/10.1016/j.acn.2005.12.002

Christ, S. E., Holt, D. D., White, D. A., & Green, L. (2007). Inhibitory control in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 37*, 1155–1165. https://doi.org/10.1007/s10803-006-0259-y

Cipolotti, L., Spanò, B., Healy, C., Tudor-Sfetea, C., Chan, E., White, M., … & Bozzali M. (2016). Inhibition processes are dissociable and lateralized in human prefrontal cortex. *Neuropsychologia, 93*, 1-12. https://doi.org/10.1016/j.neuropsychologia.2016.09.018

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological Review, 97*, 332-361. http://dx.doi.org/10.1037/0033-295X.97.3.332

Cox, C. S., Chee, E., Chase, G. A., Baumgardner, T. L., Schuerholz, L. J., Reader, M. J., Mohr, J., & Denckla, M. B. (1997). Reading proficiency affects the construct validity of the Stroop test interference score. *The Clinical Neuropsychologist, 11*, 105-110, https://doi.org/10.1080/13854049708407039

De Jong, R., Berendsen, E., & Cools, R. (1999). Goal neglect and inhibitory limitations: dissociable causes of interference effects in conflict situations. *Acta Psychologica, 101*, 379-394. http://dx.doi.org/10.1016/S0001-6918(99)00012-8

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Golden, C. J. (1978). *Stroop Color and Word Test: A manual for clinical and experimental uses.* Chicago, IL: Stoelting Co.

Golden, C. J. (1994). *Stroop: Test de colores y palabras*. Madrid, Spain: TEA Ediciones.

Golden, C. J., & Freshwater, S.M. (2002). *Stroop Color and Word Test: Revised examiner's manual.* Wood Dale, IL: Stoelting Co.

Heflin, L. H., Laluz, V., Jang, J., Ketelle, R., Miller, B. L., & Kramer, J. H. (2011). Let's inhibit our excitement: the relationships between Stroop, behavioral disinhibition, and the frontal lobes. *Neuropsychology, 25*, 655-665. doi: 10.1037/a0023863

Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences.* Oxford: Elsevier.

Kalanthroff, E., Goldfarb, L., & Henik, A. (2013). Evidence for interaction between the stop signal and the Stroop task conflict *Journal of Experimental Psychology. Human Perception and Performance, 39*, 579-592. doi: 10.1037/a0027429

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

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133,* 189–217. doi:10.1037/0096-3445.133.2.189

Kindt, M., Bierman, D., & Brosschot, 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. doi: 10.1016/0191-8869(96)00133-X

Kluttz, A., & Golden, J. C., (2016). Assessing Executive Function Measures-the Predictors of Stroop Interference T-scores. *The Clinical Neuropsychologist, 30*, 781-782. doi: 10.1080/13854046.2016.1194480

Lange, F., Kröger, B., Steinke, A., Seer, C., Dengler, R., & Kopp, B. (2016). Decomposing card-sorting performance: Effects of working memory load and age-related changes. Neuropsychology, 30(5), 579–590. https://doi.org/10.1037/neu0000271

Lanham, R. A., Vanderploeg, R. D., & Curtiss, G. (1999). The lack of construct validity of the Stroop color and word test with traumatic brain injury. *Archives of Clinical Neuropsychology, 14*, 782–783. doi: 10.1093/arclin/14.8.782

Lansbergen M. M., Kenemans J. L., & van Engeland H. (2007). Stroop interference and attention-deficit/hyperactivity disorder: A review and meta-analysis. *Neuropsychology, 21,* 251–262. doi: 10.1037/0894-4105.21.2.251

Llinàs-Reglà, J., Vilalta-Franch, J., López-Pousa, S., Calvó-Perxas, L., Rodas, D. T., & Garre-Olmo, J. (2015). The Trail Making Test: Association With Other Neuropsychological Measures and Normative Values for Adults Aged 55 Years and Older From a Spanish-Speaking Population-Based Sample. *Assessment, 24*, 1-14. doi: 10.1177/1073191115602552

Lubrini, G., Periáñez, J. A., Ríos-Lago, M., Viejo-Sobera, R., Ayesa-Arriola, R., Sánchez-Cubillo, I., … & Rodríguez-Sánchez, J. M. (2014). Clinical Spanish norms of the Stroop test for traumatic brain injury and schizophrenia. *Spanish Journal of Psychology, 17*, E96. doi: 10.1017/sjp.2014.90

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

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. doi: 10.1016/S1364-6613(00)01530-8

Mead, L. A., Mayer, A. R., Bobholz, J. A., Woodley, S. J., Cunningham, J. M., Hammeke, T. A., & Rao, S. M. (2002). Neural basis of the Stroop interference task: response competition or selective attention? *Journal of the International Neuropsychological Society, 8*, 735-742. doi: 10.1017/S1355617702860015

Penner, I. K., Kobel, M., Stöcklin, M., Weber, P., Opwis, K., & Calabrese, P. (2012). The Stroop task: comparison between the original paradigm and computerized versions in children and adults. *Clinical Neuropsychology, 26,* 1142-1153. doi: 10.1080/13854046.2012.713513

Periáñez, J. A., Maestú, F., Barceló, F., Fernández, A., Amo, C., & Ortiz Alonso, T. (2004). Spatiotemporal brain dynamics during preparatory set shifting: MEG evidence. *NeuroImage, 21,* 687–695. doi:10.1016/j.neuroimage.2003.10.008

Protopapas, A., Archonti, A., & Skaloumbakas, C. (2007). Reading ability is negatively related to Stroop interference. *Cognitive Psychology, 54*, 251-282. doi: 10.1016/j.cogpsych.2006.07.003

Regan, J. E. (1978). Involuntary automatic processing in color-naming tasks. *Perception and Psychophysics, 24,* 130-136. https://doi.org/10.3758/BF03199539

Ríos, M., Periáñez, J. A., & Muñoz-Céspedes, J. M. (2004). Attentional control and slowness of information processing after severe traumatic brain injury. *Brain Injury, 18,* 257–272. doi: 10.1080/02699050310001617442

Rosselli, M., Ardila, A., Salvatierra, J., Marquez, M., Matos, L., & Weekes, V. A. (2002). A cross-linguistic comparison of verbal fluency tests. *International*

Journal of Neuroscience, 112, 759-776. http://dx.doi.org/10.1080/00207450290025752

Rosselli, M., Ardila, A., Santisi, M. N., Arecco, M. R., Salvatierra, J., Conde, A., & Lenis, B., (2002). Stroop effect in Spanish-English bilinguals. *Journal of the International Neuropsychological Society, 8,* 819-827. http://dx.doi.org/10.1017.S1355617702860106

Sánchez-Cubillo, I., Periáñez, J. A., Adrover-Roig, D., Rodríguez-Sánchez, J. M., Ríos-Lago, M., Tirapu, J., & Barceló, F. (2009). Construct validity of the Trail Making Test: Role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *Journal of the International Neuropsychological Society, 15,* 438–450. http://dx.doi.org/10.1017/S1355617709090626.

Scarpina, F., & Tagini, S. (2017). The Stroop Color and Word Test. *Frontiers in Psychology, 8,* 557. doi: 10.3389/fpsyg.2017.00557

Shum, D. H., McFarland, K., Bain, J. D., & Humphreys, M. S. (1990). Effects of closed-head injury on attentional processes: an information-processing stage analysis. *Journal of Clinical and Experimental Neuropsychology, 12*, 247-264. doi: 10.1080/01688639008400971

Sisco, S. M., Slonena, E., Okun, M. S., Bowers, D., & Price, C. C. (2016). Parkinson's disease and the Stroop color word test: processing speed and interference algorithms. *The Clinical Neuropsychologist, 30*, 1104-1117. doi: 10.1080/13854046.2016.1188989

Spikman, J. M., Kiers, H. A., Deelman, B. G., & van Zomeren, A. H. (2001). Construct validity of concepts of attention in healthy controls and patients with CHI. *Brain and Cognition, 47*, 446-460. doi: 10.1006/brcg.2001.1320

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary (3rd ed.)*. New York: Oxford University Press.

Swick, D., & Jovanovic, J. (2002). Anterior cingulate cortex and the Stroop task: neuropsychological evidence for topographic specificity. *Neuropsychologia, 40*, 1240-1253. doi: 10.1016/S0028-3932(01)00226-3

Verbruggen, F., & Logan, G. D. (2008). Response inhibition in the stop-signal paradigm. *Trends in Cognitive Sciences, 12,* 418–424. doi:10.1016/j.tics.2008.07.005

Vitkovitch, M., Bishop, S., Dancey, C., & Richards A. (2002). Stroop interference and negative priming in patients with multiple sclerosis. *Neuropsychologia, 40,* 1570-1576. doi: 10.1016/S0028-3932(02)00022-2

Wechsler, D. (2012). *Escala de inteligencia de Wechsler para adultos – IV*. Madrid: Pearson.

Whiteside, D. M., Kealey, T., Semla, M., Luu, H., Rice, L., Basso, M. R., & Roper, B. (2016). Verbal Fluency: Language or Executive Function Measure? *Applied Neuropsychology Adult, 23,* 29-34. doi: 10.1080/23279095.2015.1004574

Table 1: *Overview of studies of relevance to Stroop construct validity*

Authors (Year)	Sample (N)	Stroop Scores	Other Cognitive Scores	Statistical Analyses	Results	Conclusions and Implications
Shum, et al., (1990)	Healthy young and middle-age controls, and traumatic brain injury patients (n=170)	DS	- Letter Cancellation - Serial Subtraction - Digit Span - DigSym - TMT-A, TMT-B - SDMT - Knox Cube	PCA	IS: loaded .59 in a Sustained Selective Processing factor (together with Serial Substraction scores), and .38 in a Visuo-motor Scanning factor (together with DigSym, Letter Cancellation, SDMT, TMT-A, and TMT-B).	Not related to Stroop test validity.
Cox, et al., (1997)	Parents of children with learning disabilities (n = 306)	IS (T score)	- WCST (Failures to maintain set, Pers. Resp., Categories) - Test of variables of attention, TOVA (Commission errors) - Word Fluency (Semantic, Phonetic) - SCN, SWR, SCW - Selective Reminding Instructions - Rey Complex Figure (Copy Organization)	Correlation	IS: correlates with TOVA Commission errors s (r > -.32), SWR (-.37), SCW (.61), and Rey Copy Organization (.34) only in subjects with high reading automaticity.	IS reflects the ability to inhibit an automatic response pattern.
Lanham, et al. (1999)	Traumatic brain injury (n=622)	SWR, SCN, SCW, IS	- Visual processing - Verbal Learning and Memory - Attention - WCST (Concept formation) - Verbal Fluency - Language - BNT (Paraphasic Errors)	Correlation PCA	SWR, SCN, and SCW loaded .62, .60 and .61 on a “Verbal Fluency” factor. IS: no significant relationship with severity of injury measures.	SWR, SCN, and SCW reflect Verbal Fluency. IS has a doubtful clinical interest.
Spikman et al., (2001)	Healthy young and old controls (n=60) (traumatic brain injury	SCW	- TMT-A, TMT-B - RT Distraction Task - PASAT-5	PCA	Healthy Controls: SCW: loaded .78 in a “Control” or Memory-driven Action component	Not related to Stroop test validity.

	n=60 not analysed here)		- RT Dual Task - 15 Words Test (LOC score) - MCST (PERSREL score)		(together with PASAT-5, TMT-B, LOC score, and PERSREL), and .45 in a "Speed" or Stimulus-driven Reaction component (together with RT Dual task, and RT Distraction task).	
Bondi et al., (2002)	Healthy control (n=51) (Probable Alzheimer's disease n=59, not analysed here)	SWR, SCN, SCW	- BNT - COWAT (Letter Fluency) - TMT-A, TMT-B - WCST (Cat, Pers. Err.) - WAIS-R (Digit span, DigSym, Vocabulary) - WISC (Block design) - WMS (delayed and immediate recall)	PCA	Healthy Controls: SWR and SCN loaded both .70 in "speeded of visual processing" factor (together with TMT A and B, Dig. Sym., WISC block design). SCW: loaded .55 in a "semantic knowledge and verbal processing speed" (together with the BNT, COWAT, Digit Span, and WAIS–R vocabulary).	SWR and SCN reflect information processing speed. SCW reflects verbal processing speed. Results also revealed a different factor structure for healthy participants and AD patients.
Ríos et al., (2004)	Traumatic brain injury patients (n=29) Healthy controls (n=30)	SCN, SWR, SCW, IS	- WCST (Pers. Err., Pers. Resp., Incorrect Resp., Correct Resp., Non-Pers. Err.) - TMT-A, TMT-B, B:A	PCA	SWR and SCN loaded .91 and .87 in a "speed factor" (together with TMT-A, TMT-B, and SCW). SCW: loaded .72 on "speed" and .58 on the "interference control" factor (together with IS). IS: loaded .98 on "interference control" factor (together with SCW).	SWR and SCN tap on 'speed of processing'. SCW taps on 'speed of processing' and 'interference control'. IS provides an indicator of 'interference control'.
Chaytor et al., (2006)	Neurological adult patients (n=46)	SCW, IS	- WCST (% Pers. Err). - TMT-A, TMT-B, B-A - COWAT	Correlations	SCW: correlates with TMT-B, B-A, IS, COWAT, WCST % Pers. Err (r > .35). IS: Only correlates with SCW (.36).	Not related to Stroop validity.
Christ et al., (2007)	Children with autism spectrum disorders (n=18) Healthy controls (23 Biological siblings of children with autism	SWR, SCN, SCW	- Stroop computerized task (RTs and errors in neutral, inhibitory trials) - Flanker task (RTs and errors in neutral, inhibitory trials) - Go-no go task (RTs and errors in go	Group differences	Patients and controls showed no differences in any Stroop score, either computerized or paper and pencil. Patients and controls differed in RTs from the inhibitory trials of the Flanker task.	Stroop, Flanker, and Go-no go tasks are measuring different aspects of inhibitory control.

	spectrum disorder, and 25 non-sibling control group.		trials and errors in no go trials)		Patients and controls differed in errors from the go trials of the go-no go task.	
Protopapas et al., (2007)	Healthy Children (n=156)	SCN, SCW, DS	- Reading skills (10 scores) - Raven's SPM - WISC-III (Arithmetic, Digit Span)	Correlation, PCA	SCN: correlates r>.39 with time in reading skills (pseudoword, word, and text reading) SCW: correlates r>.32 with time and errors in reading skills (pseudoword, word, and text reading, and with word and text spelling), and r= -.34 with Digit span. IS (SCW-SCN): correlates r>.30 with time and errors in reading skills (pseudoword, word, and text reading, and with word and text spelling). SCN and IS (SCW-SCN) loaded .51 and -.47 in a single "reading speed" factor (together with pseudowords, words, and text read times, and text read errors).	There is a direct link between reading skills and Stroop interference, beyond the effects of executive functioning.
Sánchez-Cubillo et al., (2009)	Healthy controls (n=41)	SCW	- DigSym - Finger Tapping - DFor, DBack - WCST computer version (RT Switch Cost) - TMT-A, TMT-B, B-A, B:A; Log B:A	Correlation Regression	SCW: correlates with DigSym (.48), DBack (.34), TMT-A (.34), TMT-B (.38), B-A (.31).	Not related to Stroop validity.
Heflin et al., (2011)	Patients with mild cognitive impairment or dementia (n=112)	SCN, SCW	- NPI Disinhibition	Correlation Regression	SCN: correlates -.26 with NPI Disinhibition. SCW: correlates -.22 with NPI Disinhibition. SCN predict NPI Disinhibition (r^2=.07), after controlling for SCW, Mini Mental State Examination, and age.	SCW showed no association with behavioural disinhibition. Indeed, SCN showed a stronger relationship with behavioural disinhibition than did SCW.

					Neither SCW correct nor errors do predict NPI Disinhibition (r^2=.01), after controlling for colour naming, MMSE, and age.	
Adrover-Roig et al., (2012)	Healthy middle-age and elderly controls (n=122)	SWR, SCW	- TMT-A, TMT-B, TMT B/A - DFor, DBack - Rey's Complex Figure - BNT - COWAT - Semantic Fluency - Brixton (errors) - WCST modified version (4 scores) - DigSym	Latent variable analysis (LISREL)	SWR: loaded .43 on "Speed" latent factor (together with DigSym and TMT-A). SCW: loaded .69 on "Working Memory" latent factor.	Not related to Stroop validity.
Llinas-Reglà et al, (2015)	Healthy controls (N=1923).	SCW	- Symbol Digit - Finger Tapping (MacQuarrie Test for Mechanical Ability) - DFor, DBack - TMT (Log A, Log B, Log Dif. Log Ratio) - Verbal Fluency (phonemic and semantic)	Correlations Factorial Analysis	SCW: correlated with Symbol Digit (.58) Log TMTB (-.54) Finger Tapping (.51), Log TMTA (-.46), Log TMT Dif (-.44), verbal fluency-semantic (.39), DigBack (.38), verbal fluency-phonemic (.35), DigFor (.32), Log TMT Ratio (-.16). SCW loaded > .72 in speed related factors of the four factorial analyses performed.	Not related to Stroop validity.
Sisco et al., (2016)	Patients with Parkinson Disease (N=58) and non-Parkinson disease age matched peers (N=68)	IS, DS, RelS, RatS, Residualized	- Digit symbol - Symbol Search - TMTA, TMTB, ratio - COWAT - WCST (Cat.)	Correlations	IS: did not correlate with any other cognitive score (r < .21, in all cases). DS: correlated with Digit symbol (.34), Symbol Search (.25), and TMTA (.20). RelS: correlated with TMT ratio (.27) RatS: correlated with TMT ratio (.39) and	IS did not correlate with any measures of processing speed or executive function. DS correlated significantly with standardized processing speed but not executive function

					WCST Cat (-.27). Residualized: TMT ratio (.27)	measures RelS, RatS, and Residualized scores correlated with executive function but not processing speed measures
Kluttz and Golden, (2016)	Patients with neurological disorder (50%), mood disorder (40%), and not specified (10%; n=648)	IS (T score)	- Time (T score) and errors in TMT-A and TMT-B - WCST (Trials, Pers. Err., Cat., Trial to complete 1st cat.; % conceptual level; Correct Resp; Non-Pers. Err., Total Err., Failures to Maintain Set, and Learning to Learn)	Regression	IS: only WCST total err. (β= ± .24) minimally predicted IS.	IS is related to inhibiting task irrelevant information.

Note: TMT (Trail Making Test), WAIS-R (Wechsler Adult Intelligence Test-Revised), WMS (Wechsler Memory Scale), WISC (Wechsler Intelligence Scale for Children); DigSym (WAIS-III Digit Symbol), FingT (Finger tapping), DFor (WAIS-III Digit Forward), DBack (WAIS-III Digit Backward), SCN (Stroop Colour Naming), SWR (Stroop Word reading), SCW (Stroop Colour-Word), IS (Stroop Interference score= SCW – [(SWR*SCN)/(SWR+SCN)]), DS (Stroop difference score= SCW-SCN) RelS (Stroop relative score= [(SCN-SCW)/SCN]*100), RatS (Stroop ratio score= SCW/SCN), BNT (Boston Naming Test), MMSE (Mini Mental State Examination); WCST (Wisconsin Card Sorting Test), MCST (Modified Card Sorting Test), COWAT (Controlled Word Association Test), CVLT (California Verbal Learning Test), PASAT (Paced Auditory Serial Addition Test), CPT (Continuous Performance Test), SDMT (Symbol Digit Modality Test), NPI (Neuropsychiatric Inventory); and PCA (Principal components analysis).

Table 2: *Descriptive statistics*

	Stroop Scores							Other Cognitive Measures				
	SWR	SCN	SCW	IS	DS	RatS	RelS	DigSym	Fluency A	Digit Span	TMT B-A	IncongRT
N	83	83	83	83	83	83	83	83	83	83	83	83
Mean	112.8	77.2	50.1	4.4	27.1	.65	35.1	89.2	13	16	25.9	513.4
SE	1.6	1.1	1.3	1.1	1.1	.01	1.4	1.5	.5	3	1.7	12.6
Min-Max	85-155	53-99	21-80	-19-30.6	1-48	.3-.9	1.3-68	51-126	5-28	7-24	-19-69	327-943

Note: SWR (Stroop Word-Reading); SCN (Stroop Colour-Naming); SCW (Stroop Colour-Word); IS (Stroop Interference score= SCW – [(SWR*SCN)/(SWR+SCN)]); DS (Stroop difference score= SCN-SCW); RatS (Stroop ratio score= SCW/SCN); RelS (Stroop relative score= [(SCN-SCW)/SCN]*100; DigSym (WAIS-IV Digit Symbol), Fluency A (Phonemic Verbal Fluency task with the letter A); Digit Span (WAIS-IV Digit Forward + Digit Backward), TMT B-A (Trail Making Test derived B-A score), IncongRT (Reaction Time in the incongruent condition of a computerized Stroop task), and SE (Standard error of measurement).

Table 3: *Correlation matrix*

	SWR	SCN	SCW	IS	DS	RatS	RelS	DigSym	Fluency A	Digit Span	TMT B-A	IncongRT
SWR	1											
SCN	.55**	1										
SCW	.38**	.55**	1									
IS	.01	.16	.90**	1								
DS	.10	.35**	-.59**	-.85**	1							
RatS	.10	.01	.83**	.96**	-.93**	1						
RelS	-.10	-.01	-.83**	-.96**	.93**	-1**	1					
DigSym	.22*	.27*	.31**	.22*	-.09	.22	-.22	1				
Fluency A	.11	.15	.21 (*)	.17	-.09	.17	-.17	.22*	1			
Digit Span	.10	.28**	.31**	.24*	-.07	.20	.20	.01	.28*	1		
TMT B-A	-.16	-.08	-.13	-.08	.07	-.10	.10	-.15	-.13	-.12	1	
IncongRT	-.11	-.26*	-.33**	-.27*	.12	-.25*	.25*	-.33**	-.34**	-.08	.23*	1

Note: ** p< 0.01; * p< 0.05; (*). p< 0.06 (Two-tailed); SWR (Stroop Word-Reading); SCN (Stroop Colour-Naming); SCW (Stroop Colour-Word); IS (Stroop Interference score = SCW – [(SWR*SCN)/(SWR+SCN)]); DS (Stroop difference score= SCN-SCW); RatS (Stroop ratio score= SCW/SCN); RelS (Stroop relative score= [(SCN-SCW)/SCN]*100; DigSym (WAIS-IV Digit Symbol), Fluency A (Phonemic Verbal Fluency task with the letter A); Digit Span (WAIS-IV Digit Forward + Digit Backward), TMT B-A (Trail Making Test derived B-A score), IncongRT (Reaction Time in the incongruent condition of a computerized Stroop task).

Table 4: *Results of multiple regression analysis on Stroop direct and derived scores.*

	β	*t*	*p*	Zero-Order	Partial
SWR					
DigSym *	.221	2.04	.045	.221	.221
Fluency A	.067	0.6	.550	.113	.067
Digit Span	.106	0.98	.331	.109	.109
TMT B-A	-.129	-1.18	.242	-.160	-.131
IncongRT	-.047	-0.4	.688	-.114	-.045
SCN					
DigSym *	.270	2.63	.010	.274	.282
Fluency A	.011	0.1	.921	.148	.011
Digit Span *	.280	2.71	.008	.283	.291
TMT B-A	-.003	-0.03	.979	-.076	-.003
IncongRT	-.167	-1.54	.128	-.259	-.171
SCW					
DigSym *	.223	2.25	.027	.314	.245
Fluency A	-.001	-0.01	.993	.210	-.001
Digit Span *	.289	2.94	.004	.311	.314
TMT B-A	-.002	-0.02	.982	-.125	-.003
IncongRT *	-.234	-2.25	.027	-.334	-.245
IS					
DigSym	.145	1.31	.192	.217	.146
Fluency A	.026	0.23	.822	.169	.025
Digit Span *	.219	2.09	.040	.239	.227
TMT B-A	.004	0.04	.967	-.080	.005
IncongRT *	-.257	-2.45	.017	-.274	-.264
RatS					
DigSym	.151	1.33	.188	.215	.147
Fluency A	.095	0.83	.407	.167	.093
Digit Span	.185	1.73	.087	.203	.190
TMT B-A	-.049	-0.44	.662	-.103	-.049
IncongRT *	-.245	-2.27	.026	-.245	-.245

Note: SWR (Stroop Word-Reading); SCN (Stroop Colour-Naming); SCW (Stroop Colour-Word); IS (Stroop Interference score= SCW – [(SWR*SCN)/(SWR+SCN)]); RatS (Stroop ratio score= CW/CN); DigSym (WAIS-IV Digit Symbol), Digit Span (WAIS-IV Digit Forward + Digit Backward); Fluency A (Phonemic Verbal Fluency task with the letter A); IncongRT (Reaction Time in the incongruent condition of a computerized Stroop task); *= variables showing significant contributions to the regression model.

STROBE Statement—checklist of items that should be included in reports of observational studies

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported
Objectives	3	State specific objectives, including any prespecified hypotheses
Methods		
Study design	4	Present key elements of study design early in the paper
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
Participants	6	(*a*) *Cohort study*—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study*—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants
		(*b*) *Cohort study*—For matched studies, give matching criteria and number of exposed and unexposed *Case-control study*—For matched studies, give matching criteria and the number of controls per case
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group
Bias	9	Describe any efforts to address potential sources of bias
Study size	10	Explain how the study size was arrived at
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding
		(*b*) Describe any methods used to examine subgroups and interactions
		(*c*) Explain how missing data were addressed
		(*d*) *Cohort study*—If applicable, explain how loss to follow-up was addressed *Case-control study*—If applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy
		(*e*) Describe any sensitivity analyses

Continued on next page

Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed
		(b) Give reasons for non-participation at each stage
		(c) Consider use of a flow diagram
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders
		(b) Indicate number of participants with missing data for each variable of interest
		(c) *Cohort study*—Summarise follow-up time (eg, average and total amount)
Outcome data	15*	*Cohort study*—Report numbers of outcome events or summary measures over time
		Case-control study—Report numbers in each exposure category, or summary measures of exposure
		Cross-sectional study—Report numbers of outcome events or summary measures
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included
		(*b*) Report category boundaries when continuous variables were categorized
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses
Discussion		
Key results	18	Summarise key results with reference to study objectives
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
Generalisability	21	Discuss the generalisability (external validity) of the study results
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.