Test-Retest Reliability of Four Cognitive Tasks in a not so WEIRD Sample

Jose A. Rodas[a,b*], Ciara M. Greene[a]

[a] *School of Psychology, University College Dublin, Ireland*
[b] *Facultad de Ciencias Psicológicas, Universidad de Guayaquil, Ecuador*

Abstract

Most cognitive tasks have been designed and validated in Western, Educated, Industrialised, Rich, and Democratic (WEIRD) populations, raising questions about the generalisation of their psychometric properties to culturally distinct contexts. This study assessed the test-retest reliability of four widely used cognitive tasks—the Colour-Word Stroop Task, the Sustained Attention to Response Task (SART), the Stop-Signal Task, and the Updating Letter Memory Task—in an Ecuadorian sample of university students. Data were collected across two sessions separated by one week and analysed using Pearson and Intraclass Correlation Coefficients (ICCs). Results indicate that most tasks exhibit reliability levels comparable to those reported in WEIRD populations, with scores ranging from moderate to good. However, limitations in the reliability of derived measures, such as the Stroop inhibition index, were identified, potentially influenced by cultural or methodological factors. This study suggests that the cognitive tasks evaluated are largely reliable in non-WEIRD cultural contexts, supporting their cross-cultural applicability. The findings also highlight the importance of assessing reliability in diverse populations to enhance the ecological validity of these tools.

Keywords: *reliability, Stroop, stop-signal, WEIRD population, SART, letter memory.*

It has been observed that most of the research in psychology published in scientific journals utilizes samples from Western, Educated, Industrialised, Rich, and Democratic (WEIRD) societies (Henrich et al., 2010), which represent only a very small and specific population in relation to the rest of the world. Additionally, many of these studies use university students as samples because they are usually easier to reach. Considering that in many rich western countries university fees tend to be high, these samples present even more specific characteristics. This type of population is not only educated but also has the economic, intellectual or personality resources (in order to gain access to a scholarship, for example) to register at a university. The specific characteristics of this population have generated some concerns among researchers about the extent to which these results can be generalized to other populations (Rad et al., 2018), particularly when considering the great cultural variability existing between countries and that these cultural differences can have important psychological effects in individuals (Oyserman & Lee, 2008). Even though many of the variations found between samples of different countries could be attributed to social and cultural factors, shared characteristics within non-WEIRD samples could also have effects on cognition, for example, sustained poverty can lead to problems in the neurological development of children (Duncan & Brooks-Gunn, 2000; McLoyd, 1998), which can result in cognitive deficiencies. Other cognitive differences derived from culture have also been observed; for example, Masuda and Nisbett (2001) found that Japanese participants paid more attention to the contextual features of the stimuli than American participants. Additional cognitive differences have also been observed between East Asian and Western cultures (Nisbett, 2003) raising questions regarding the stability of cognitive processes, and of the psychometric properties of the instruments used for their measurement, across national and cultural boundaries.

Despite these observations there is still the notion among psychologists that these differences might not be as important or deep as they may seem (as seen in the level of generalisation given to results of many published studies). This may be especially true in the field of cognitive psychology, where basic functions such as attention, memory, and perception are studied. In a commentary on Henrich et al.'s (2010) article, Gaertner and others (2010) argue that the level of analysis plays a key role when determining if cultural differences can lead

*Correspondence concerning this article should be addressed to Jose A. Rodas. E-mail: jose.rodas@ucd.ie
Received 16 August 2024; Accepted 28 October 2025; Available online 9 December 2025.

to discrepancies in psychological phenomena or behaviour across populations. Gaertner et al. present dietary behaviours as an example: members of different cultures frequently follow specific diets (e.g. excluding the meat of a particular animal) but when these behaviours are analysed at a more basic level, we observe that all humans require similar levels of sustenance and nutrition. Similarly, although the basic function of attention is observable across humankind, specific ways in which an attentional process is utilised – e.g. in social situations - may be distinct from one population to the other. One limitation to this approach is that it is not clear how deeply culture can affect cognition and its processes.

A growing body of research has examined the stability and capacity of cognitive tasks to measure the processes they were designed to evaluate, particularly in underrepresented populations like those in Latin America (LA). Executive functions, which include attentional control, working memory, inhibition, and cognitive flexibility, are critical for goal-directed behaviour (Lezak et al., 2012; Stuss, 2011). These functions develop during childhood and adolescence alongside brain maturation, particularly in the frontal lobes (Petersen & Posner, 2012). Studies such as Dias et al. (2015) demonstrate the cross-cultural applicability of EF frameworks. For example, the tripartite model of EF proposed by Miyake et al. (2000)—comprising working memory, inhibition, and cognitive flexibility—was replicated in a Brazilian sample, highlighting the need for cognitive tasks to maintain stability and validity across diverse cultural contexts.

However, many studies from LA are published in Spanish or Portuguese, limiting their accessibility to the global academic community. This is particularly problematic for normative data on EF tasks in non-WEIRD populations. Research by Fierro Bósquez et al. (2024) and Rivera et al. (2015) on EF and attentional tasks in LA populations highlights both universal cognitive patterns and cultural influences. Contextual biases, such as those observed by Masuda and Nisbett (2001) in Japanese participants, may similarly affect task performance in LA populations. These findings emphasise the importance of rigorous psychometric testing to ensure reliability and validity across cultures. This study examines the test-retest reliability of four widely used EF tasks—the Colour-Word Stroop task, the Sustained Attention to Response Task (SART), the Updating Letter Memory Task, and the Stop-Signal task—in an Ecuadorian sample. This population provides a unique context, differing substantially from WEIRD populations in cultural, linguistic, and socioeconomic characteristics. By addressing these gaps, we aim to advance the understanding of cognitive task stability and their cross-cultural applicability.

The role of socio-cultural factors in shaping cognitive processes such as inhibition, updating, and sustained attention remains underexplored, particularly in underrepresented populations. Foundational functions like attention are often presumed to be universal due to their centrality to information processing, yet evidence suggests they can be culturally influenced. Nisbett (2003) and Masuda and Nisbett (2001) documented significant differences in attentional focus and cognitive styles between Western and East Asian cultures. Similarly, research from LA demonstrates how cultural and linguistic contexts affect performance of EF measures (Fierro Bósquez et al., 2024; Rivera et al., 2015).

These findings highlight the need to examine how socio-cultural factors influence EF task performance, particularly in non-WEIRD populations. Tasks developed in Western contexts may not account for culturally specific cognitive patterns or biases, underscoring the necessity of psychometrically robust and culturally relevant assessments for diverse populations.

A not so WEIRD population

LA countries represent good cases of samples with very different characteristics than the WEIRD samples commonly used. The mean number of years of education tend to be closer to the mean of all countries in the world, problems with democratic procedures are commonly reported, income per capita tends to be very low, and populations tend to have a more variable demographic background. The present study was conducted in Ecuador, where 72% of the population identify themselves as a mixture of native American, African and European populations, 23% live on less than $85 a month, which is half of the income considered necessary for one person to live in the country (Instituto Nacional de Estadística y Censos, 2018), and the mean number of years of education in the population is 8.7 compared to 13.4 in the United States, 12.9 in the United Kingdom, and 8.4 globally (United Nations Development Programme, 2018). According to the 2018 United Nations Development programme report, the gross national income per capita in Ecuador was $10,347[1] in 2017, whereas the mean world

[1] 2011 PPP dollars were used in the report. PPP stands for Purchasing Power Parity and is a common measure in macroeconomics to compare the purchasing power of a country with other countries as it not only reflects the exchange rates but also compares the cost of a common basket of goods between countries. It is based on the purchasing power of a dollar within the United States during a specific period of time, in this case the year 2011.

income per capita was $15,295 and $10,055 for developing countries in the same year. In comparison, the income per capita in the United States and the United Kingdom in 2017 was $54,941 and $39.116, respectively (United Nations Development Programme, 2018). For a more in-depth analysis of the social and cultural aspects of the Ecuadorian population, see Capella and others (2019).

The sample in this study consists of Psychology undergraduate students from the University of Guayaquil, a public university in Ecuador where public education is offered for free. Ecuadorian law specifically precludes the university from charging any fees to students, and the university cannot oblige them to incur into any expense directly related to their academic obligations. Guayaquil is the biggest and most economically active city in Ecuador, a situation that has contributed to making the University of Guayaquil the biggest in the country with more than 60 thousand students from different regions of Ecuador by 2016 (Universidad de Guayaquil, 2016). Another important aspect of this sample is its limited prior contact with the cognitive tasks used in this study. Although psychological research is conducted in the university, students have very limited exposure to it, as Psychology education is entirely focussed on professional practice (Capella & Andrade, 2017). This has had several effects on the psychology degree offered in the university; for example, research-focused fields such as cognitive psychology do not form part of the curriculum. The participants in the present study were therefore completely naïve to the methods and procedures used in this research.

Test-retest reliability

Reliability is one of the two main psychometric properties of a psychological instrument and can be divided into internal and external reliability. Internal reliability is understood as the internal consistency of an instrument, that is, how well each item of the instrument relates to the other items that measure the same variable. Probably the most widely used method in its estimation is Cronbach's Alpha (Cronbach, 1951; Tavakol & Dennick, 2011). External reliability refers to how stable the scores obtained from different measurements are over a short period of time. Even though both forms of reliability measure different psychometric properties of an instrument, it is common to find only Cronbach's Alpha reported in reliability studies. External reliability has been recommended over internal consistency by several authors, in particular, the test-retest method (Leppink & Pérez-Fuster, 2017; McCrae et al., 2011). This is particularly important for cognitive tasks, as in most cases the items within each task (trials) are very similar if not identical to one another, which makes internal consistency less relevant than the stability of scores across testing sessions.

As the name implies, the test-retest method requires the administration of the task on at least two different occasions within a short period of time, after which those scores are compared. Estimating the interclass correlation coefficient (Pearson's *r*) is one of the most common procedures for comparing the scores as it reflects their correlation. One drawback of using only Pearson's *r* is that it cannot detect systematic differences or biases between scores. For example, if scores from two different measurements, A and B, are compared it is possible that scores from measurement B may be consistently and significantly higher than scores from measurement A, while still producing a very high *r* value. For this reason, it often proves convenient to test for significant differences between the two measurements by using hypothesis tests such as *t*-tests or ANOVAs. This latter type of analysis also allows for an agreement inspection between scores - that is, an assessment of the instrument's ability to produce the exact same score on two occasions if no change has occurred in the object being measured (Berchtold, 2016). An alternative method has been suggested for measuring test-retest reliability that allows for a correlation and agreement inspection within a single score: the intraclass correlation coefficient (ICC; Mcgraw & Wong, 1996). An important difference between the inter and intraclass correlations is that the interclass correlation (i.e. Pearson's *r*) allows comparison between measurements of different "classes", or types of data (e.g. weight or the number of kilometres the person can run), whereas the intraclass correlation only allows for comparisons between elements of the same class, as in the case of scores from different raters using the same instrument (McGraw & Wong, 1996). Thus, the ICC is the most appropriate measure for assessing test-retest reliability as it provides a single score ranging from 0 (no reliability at all) to 1 (superb reliability) that includes both the correlation and agreement between measurements.

Tasks included in the present study

The four tasks selected for the current study are commonly used when investigating the cognitive processes of inhibition, updating, and sustained attention in experimental and clinical settings. These are key processes for adequate cognitive functioning and behaviour. Inhibition, for example, regulates behaviour by controlling impulses considered inappropriate in a given situation (Diamond, 2013). The ability to update the contents of working memory (henceforth referred to

simply as 'updating') is critical to information processing, as the capacity of WM is very limited, and information needs to be continually replaced (Kessler & Oberauer, 2014). Finally, the capacity to sustain attention on a task is essential for many of the activities of daily living, such as driving or grocery shopping, and tends to be affected in several psychopathologies (Brands et al., 2005; Ebert & Kohnert, 2011).

One other important reason for selecting these tasks, in particular, is because their psychometric properties in "WEIRD" samples have been reported in numerous articles (for a detailed description of each task consult the Methods section). The first of these tasks, the Stop-Signal Task, was developed to evaluate inhibitory control of an impulse (Logan et al., 1997). The most important value produced by this task is the stop-signal reaction time (SSRT), a measure of the speed with which participants can inhibit a pre-potent response once a stop signal is presented. Several laboratories in Europe (Bekker et al., 2005; De Zeeuw et al., 2008), the United States (Blaskey, Harris, & Nigg, 2008), and Canada (Toplak et al., 2009) have shown inhibitory control differences between ADHD patients and controls using this task. Test-retest reliability has been evaluated with ADHD and other clinical samples from Canada and United States (Soreni et al., 2009; Weafer et al., 2013), finding moderate reliability (see Table 1). Laboratories in Europe have also evaluated its test-retest reliability, although some of them have shown mixed results.

Table 1

Reliability scores from previous studies using WEIRD samples

Study	Origin of sample[a]	ICC	Pearson's *r*
Colour Word Stroop (inhibition score)			
Hedge et al. (2018)[b]	United Kingdom	0.60	
Hedge et al. (2018)[c]	United Kingdom	0.66	
Siegrist (1995) [b]	Switzerland		0.73
Siegrist (1997) [c]	Switzerland		0.68
Strauss et al. (2005)	United States		0.46
SART (errors of commission)			
Robertson et al. (1997)	United Kingdom		0.76
Stop-Signal Task (SSRT)			
Hedge et al. (2018) [b]	United Kingdom	0.47	
Hedge et al. (2018) [c]	United Kingdom	0.43	
Kuntsi et al. (2001)	United Kingdom	0.11	
Soreni et al. (2009)	Canada	0.72	
Weafer et al. (2013)	United States		0.65
Wöstmann et al. (2013)	Germany	0.03	0.03

Note. ICC = Intraclass Correlation Coefficient

[a] This column indicates the country in which the sample was obtained. Unfortunately, the included studies did not report nationality of the members of the sample, and we therefore cannot preclude the possibility that participants from non-WEIRD countries were included.

[b] Study 1

[c] Study 2

For example, in a study from the UK and the Netherlands investigating the reliability of several tasks, extremely low reliability was reported for the SSRT (Kuntsi et al., 2001), and another study from the UK and Germany (Wöstmann et al., 2013) reported no reliability at all (i.e. a non-significant interclass correlation). A search in LA journals from the Scopus and Redalyc databases using the term "Stop Signal" revealed that, although the task is mentioned in more than 50 studies, only four used it within a LA population, and none reported any psychometric properties.

The Colour-Word Stroop task is also a common measure of inhibitory control, although it was not originally designed for that purpose (Stroop, 1935). Participants are required to respond to the colour in which a word is printed while ignoring the meaning of the word; for example, if the word "red" is printed in blue ink, participants should respond "blue", thus inhibiting their automatic response to the word's semantic content. The task has also been used as a measure of selective attention or attentional bias (Atkinson et al., 2009; Epp et al., 2012) and processing speed (Van Den Heuvel et al., 2006). It has been validated as an executive function task in different ways, for example, it has been observed in several neuroimaging studies that the anterior cingulate cortex, an area of the brain thought to be critical for selective attention, plays a prominent role during Stroop performance (Botvinick et al., 2004; Pardo et al., 1990). The traditional Stroop task was administered using printed cards; however most current studies use a computerised version which can record the latency of keypress responses with a high degree of precision (Dalgleish, 1995).

In a study with an American sample, good test-retest reliability scores have been found with the card format of the task (Franzen et al., 1987). Test-retest reliability has also been tested with Swiss and American populations using the computerised version of the task, and good reliability has also been found (Siegrist, 1995, 1997; Strauss et al., 2005). Hedge and others (2018) found moderate reliability scores in a UK sample across two different studies using the computerised version of the task. Several studies have evaluated the test-retest reliability of the card version of the task in LA samples (Rodríguez Barreto et al., 2016), finding in general good reliability scores, however, to our knowledge no reliability analyses of the computerised version have been conducted in this population. The card and computer versions of the task register responses very differently: in the card version of the task the total time taken to read all stimuli of one type is recorded, while in the computer version the common procedure is to record the response time (RT) to each stimulus presented. These different formats could affect the reliability of the task, and it is, therefore, important to assess test-retest reliability in this population using the now ubiquitous computerised task.

The SART was developed as a measure of sustained attention and requires participants to respond to a continuous stream of stimuli while withholding response to a rare target. The original authors (Robertson et al., 1997) provided evidence of its validity and reliability with a British sample. In this article, the authors observed a significant correlation between the SART and other tests of sustained attention and performed a test-retest correlation finding adequate reliability. A search for studies assessing the psychometric properties of this task with LA population was performed using the Redalyc and SciELO databases, however no studies reporting reliability could be found[2]. Cheyne and others (2009) have proposed that more specific attentional problems - namely focal inattention, global inattention and behavioural disengagement - may be evaluated using additional scores obtained in the SART as proxies of task engagement difficulties. These three attentional problems occur in the form of progressive stages, where during the first stage participants disengage slightly from the task, causing their performance to fluctuate (e.g. if they mind-wander for a very short time their response latency increases), during the second stage, attention to the task becomes unstable and, finally, in the third stage, the participant's attention is completely off-task, and they do not respond at all. One limitation of these measures is that to our knowledge, no test-retest reliability estimates have been published.

The Updating Letter Memory task (UMLT) (Miyake et al., 2000) was adapted from a recognition task (Morris & Jones, 1990) and modified in such a way that it allowed for evaluation of updating and monitoring of information held in working memory. Although other methods have also been used to evaluate this function (e.g. Garavan, 1998; Kessler & Oberauer, 2014) the ULMT has shown strong evidence of validity (assessed through confirmatory factor analysis with other tasks expected to evaluate updating; Friedman et al., 2008; Miyake &

[2]The terms 'SART', 'sustained attention' and 'atención sostenida' were used as search strings.

Friedman, 2012). The updating process has not been investigated as thoroughly as the other tasks (Ecker et al., 2010; Kessler & Oberauer, 2014), and there is little information about the reliability of instruments used to assess it, including the ULMT. Nevertheless, the original authors have reported the internal consistency of the task on several occasions (Friedman et al., 2006; Miyake et al., 2000), finding it adequate. In a study with a sample of Brazilian participants who had been diagnosed with schizophrenia, Berberian and others (2015) investigated the psychometric properties of the ULMT. It adequately correlated with another task requiring the updating of information from different categories and presented good internal consistency. Unfortunately, test-retest reliability was not investigated.

Methodology

Study Design

This study employed a test-retest design to assess the reliability of four widely used cognitive tasks—the Colour-Word Stroop Task, Sustained Attention to Response Task (SART), Stop-Signal Task, and Updating Letter Memory Task—in a sample of psychology undergraduate students from the University of Guayaquil, Ecuador. Participants were tested in group settings with computerised versions of these tasks, which measure cognitive processes such as inhibition, sustained attention, and working memory updating. Data were collected across two sessions separated by one week, and reliability was evaluated using Intraclass Correlation Coefficients (ICCs) to assess score stability. The design aimed to examine whether cultural and demographic differences in this non-WEIRD population influenced the tasks' psychometric properties.

Participants

The study was advertised via flyers in the Faculty of Psychology at the University of Guayaquil as a cognitive psychology study investigating mental processes through experimental tasks. A total of 185 undergraduate students (111 female, mean age = 19.54, SD = 3.53) responded to the advertisement and participated in an initial testing session. All participants were expected to complete all tasks; however, not all participants agreed to complete all tasks, and technical issues with the computers used for data collection led to the loss of some data. Consequently, 96 participants completed the retest of the Colour Word Stroop, 88 the SART, 86 the Updating Letter Memory Task, and 115 the Stop-Signal task. Some participants also expressed difficulties in understanding the task instructions, and their data were excluded from analysis. The final sample sizes for analysis were 94 for the Colour Word Stroop, 87 for the SART, 86 for the Updating Letter Memory Task, and 112 for the Stop-Signal Task. No specific inclusion criteria were applied; all students who volunteered and provided informed consent were allowed to participate.

Instruments

Colour Word Stroop (Stroop, 1935)

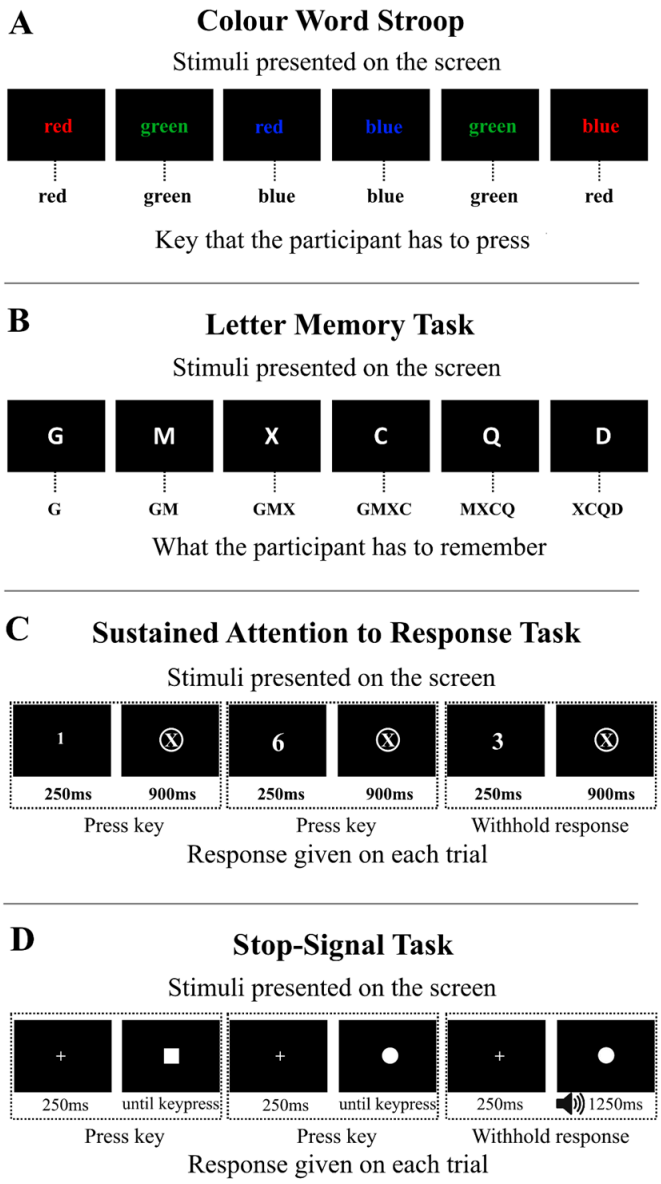
In this task the words "red", "green" or "blue" were presented one at a time in the centre of a screen, and the font colour of each word was either congruent with the word (e.g. the word "red" in a red font) or incongruent (e.g. the word "blue" in a green font). The participant was instructed to press a specific key in response to the font colour of the word while ignoring the meaning of the word (see Figure 1A). The task included 48 congruent and 24 incongruent trials, in addition to 18 congruent and 9 incongruent practice trials. The proportion of congruent and incongruent trials was similar to that used by Kane and Engle (2003), where the proportion of congruent trials was increased relative to the incongruent trials in order to strengthen the Stroop effect. Only congruent and incongruent conditions were used, and both types of stimuli were presented within a single block in a pseudorandom order. The outcome measure most commonly used in this task is the difference in RT between incongruent and congruent trials (MacLeod, 1991), which represents the inhibition score.

Sustained Attention to Response Task (Robertson et al., 1997)

Single white digits of different font sizes briefly appeared on a black screen (250ms) followed by a mask consisting of a white X inside a circle (900ms). The presentation of each digit and its corresponding mask is considered a trial. Participants were instructed to press the space bar each time any digit other than 3 appears on the screen (go trials) and to withhold the response when the digit 3 is presented (no-go trials; see Figure 1B). The digits used in the task were the numbers from 1 to 9 and are presented in pseudorandom order. The font sizes (48, 72, 94, 100, and 120 pixels height) were equally distributed across the nine digits. Participants first completed 27 practice trials consisting of three presentations of each digit. After the practice trials, participants completed 225 trials consisting of 25 presentations of each digit. Several outcome measures can be obtained from the task, although the number of errors by commission, produced by pressing the key on a no-go trial, is the most commonly used. In relation to the three progressive stages of task disengagement proposed by

Cheyne et al. (2009): (1) focal task inattention was measured by estimating the reaction time coefficient of variability (RTCV), calculated by dividing the standard deviation of the response time on go trials by the mean response time on go trials; (2) global task inattention was measured by the number of anticipated responses, that is, responses given in the first 200ms after the stimulus was presented; and (3) response disengagement was measured by the number of errors of omission (not pressing the key on go trials).

Figure 1



Note. Stimulus presentation in the four tasks. Each black square represents what is presented on the screen at a given moment. In panel A the words 'red', 'green', and 'blue' are presented only as an example, as in the real task used with the Ecuadorian sample the Spanish translation of these words were used ('rojo', 'verde', and 'azul' respectively). The groups of black squares enclosed in dotted lines in panels C and D represent a trial.

Updating Letter Memory task

This task was adapted from Miyake et al. (2000) and is commonly used for evaluating working memory updating performance. Upper case letters were presented centre-screen one at a time in a continuous series. Participants were instructed to keep track of the last four letters presented. The number of letters presented in each list could be either 5, 7, 9, or 11 and was unknown to the participant. For this reason, the participant was required to continuously update the information held in working memory, remembering the new letter presented and forgetting any letters presented prior to the last four (see Figure 1C for an example). After all letters from a list had been shown, the participant was asked to type the last four letters in the correct order of presentation. The lists of letters were randomly generated at the start of the task, and only included consonants. The letters "B" and "W" were not included, as the letter "B" is phonetically similar to another letter and the letter "W" is phonetically longer than the rest of the letters. Four lists were presented as practice with their length chosen randomly. After the practice trials were completed, four lists of each length (16 lists in total) were presented in random order as the experimental trials. Each letter remained onscreen for 2 seconds, and the interstimulus interval was 500ms. The outcome measure was the number of letters recalled correctly.

Stop-Signal Task (Logan, 1994)

For this study, the program STOP-IT was used to collect the data and the program ANALYZE-IT to process it in order to obtain the outcome measures, both designed by Verbruggen and others (2008). The authors have made both programs openly available for download on the Open Science Framework (https://osf.io/wuhpv/). Each trial of this task started with the presentation of a fixation sign ("+") in the centre of the screen for 250ms followed by the primary-task stimulus which was either a square or a circle. The primary-task stimulus remained onscreen for 1250ms or until the participant pressed the corresponding key, which was "z" for the square and "/" for the circle. Participants were instructed to press the corresponding key as quickly and accurately as possible when the stimulus was presented. Twenty-five per cent of the time, the stimulus was followed by a stop-signal - in this case, is a 750 Hz sound, presented for 75 ms – which indicated that participants should inhibit their response and avoid pressing any key. The sound was always presented with a delay in relation to the presentation of the stimulus; this delay, referred to as the Stop Signal Delay (SSD) started at 250ms. It then increased by 50ms after each successful response inhibition and decreases 50ms after each failure to inhibit the response. The interstimulus interval was 2 seconds, and 224 trials were presented in total. See Figure 1D for some sample trials. The trials were grouped in four blocks, the first block consisted of 32 practice trials, and the other three were the experimental blocks and included 64 trials each. The most common outcome measure used is the stop-signal reaction time (SSRT) which is calculated by subtracting the mean SSD from the mean response time on trials without the stop-signal and is a measure of the internal stop process of a response (for more details about the model see Logan, 1994).

All tasks were computer-administered and presented with a screen resolution of 1920 x 1080 pixels. All computers used USB keyboards. The Colour Word Stroop, SART and ULMT were programmed using PsychoPy2 (Peirce, 2008).

Procedure

Participants were invited to sign up to one of eight group testing sessions. These sessions took place in a quiet room in the faculty of Psychology containing 30 computers. The number of participants per group ranged from 25 to 30, and all sessions took place within a single week. At the beginning of each session participants were briefed about the experiment, explaining that a retest was going to take place the following week, and were asked to sign the consent form. After the participants agreed to participate and signed the consent form the assessment started. The order of task presentation was counterbalanced using a Latin Square. Each task started with the on-screen presentation of the instructions and participants were advised to read them in silence while the researcher read them aloud to the entire group. For the Stop-signal task, the instructions were printed on a sheet of paper in Spanish and handed to each participant at the beginning of the testing session as the program used for administering the task presents the instructions in English. During task administration, all participants used headphones as the Stop-signal task emits sounds. In addition to the four tasks reported in this article, participants also completed another task which has not been described in this article, as it was part of a different study.

Due to space and resource constraints, separate retest sessions were held for each task and were completed one week after the initial session. The retest session took place in groups of 30 participants in the same room as the initial testing session.

Ethics and Consent to Participate

The study was approved by the ethics committee of the faculty of Psychology from the University of

Guayaquil, and participants did not receive any compensation for participating. Before participation, the aims of the study were presented and an informed conscent was signed. All participants were informed that the data would be anonymised and used for research purposes.

Data Analysis

The data analysis was performed to assess the test-retest reliability of four cognitive tasks using the Intraclass Correlation Coefficient (ICC), following the two-way mixed-effects model for absolute agreement (Shrout & Fleiss, 1979). Prior the reliability analyses, Shapiro-Wilk test were performed and distribution plots were created to analyse whether data followed a normal distribution. The ICC values were interpreted based on Koo and Li's (2016) guidelines, categorising reliability as poor (<0.5), moderate (0.5–0.75), good (0.75–0.9), or excellent (>0.9). Additionally, Pearson's correlation coefficients (r) were calculated for comparison with previous studies. The analysis included point estimates and 95% confidence intervals for ICC scores but excluded p-values, as recommended for reliability assessments (McGraw & Wong, 1996). Descriptive statistics were reported for task performance at two time points, and ICCs were computed for primary measures from each task, including reaction times, error rates, and derived scores such as the inhibition index for the Stroop task. The results were summarised to facilitate cross-cultural comparisons with reliability estimates from WEIRD populations.

Results

The data and Supplementary Materials for this study may be found online at https://osf.io/da39c/. Results from the Shapiro-Wilk tests indicated that most variables significantly deviated from a normal distribution based on their p-values. However, the majority of W scores exceeded .9, and the distribution plots showed no substantial deviations from normality. This suggests that, while the data for most variables do not strictly adhere to a normal distribution, the deviations are unlikely to introduce bias in the results when using ICC and Pearson correlation coefficients. Details of the W scores and distribution plots are available in the Supplementary Materials available online.

Analyses for the Colour Word Stroop were performed with data from 94 participants (60 female, mean age = 19.65, SD = 3.83). As can be observed in Table 2, the ICCs of congruent and incongruent trials are moderate, with a range from moderate to good included in the 95% confidence interval. The interference score, obtained from the difference between response times of both type of trials, presents poor reliability, ranging from poor to moderate within the 95% confidence interval. The analyses for the SART were performed with data from 87 participants (55 female, M age = 19.62, SD = 3.96) and the ICC of the number of correct responses, number of errors by commission, mean response time on go trials, and anticipatory responses are moderate to good in both point estimates of the ICC and 95% confidence intervals. The ICCs of the number of errors by omissions and RTCV are moderate, and poor to moderate based on the 95% confidence interval.

Analyses for the ULMT were performed with data from 86 participants (56 female, M age = 19.33, SD = 2.21). As observed in Table 2, the ICC for the number of letters recalled correctly is good, and the 95% confidence interval ranged from moderate to good. Finally, ICC scores for the Stop-signal task were performed with data from 112 participants (69 female, M age = 19.54, SD = 3.56). As shown in Table 2, the ICCs for the probability of responding on stop-signal trials, stop-signal delay, and mean no-signal response time are good, and with a 95% confidence interval the ICCs are moderate to good. In the case of the stop-signal reaction time the ICC is moderate, and with a 95% confidence interval is moderate to good.

Discussion

It has been observed that cognitive differences can exist between cultures (Nisbett, 2003) and that these differences can have a specific effect on the performance of cognitive tasks depending on the culture (Masuda & Nisbett, 2001). This raises the question of how reliable these cognitive tasks, used in research and clinical settings, really are across populations with different cultural backgrounds. The comparison of our data with those observed in different test-retest reliability studies suggests that reliability appears to be less affected by cultural variables than by other types of variables, such as the specific design of the task, the instructions given to the participants, or by uncontrolled variables in a sample (Wöstmann et al., 2013). Support for this idea is found in the great variability observed in the reliability scores observed across different studies in populations with very similar cultures (see Table 1). For example, the ICC point estimate of the inhibition score of the Stop-Signal Task (SSRT) ranges from 0.03 to 0.72, with the Ecuadorian score displaying the second highest score. This means that the SSRT values observed in the Ecuadorian sample are as reliable as those observed in other populations. The other task that allows for this type of comparison, due to the number of studies reporting test-retest reliability, is the Colour Word Stroop.

Table 2

Summary of test-retest reliability scores from the Ecuadorian sample

	N	Time 1		Time 2		Pearson's *r*	Intraclass Correlation	95% Confidence Interval	
		M	*SD*	*M*	*SD*			Lower Limit	Upper Limit
Colour Word Stroop									
Congruent	94	0.82	0.23	0.74	0.18	0.59**	0.69	0.50	0.81
Incongruent	94	0.95	0.31	0.82	0.21	0.69**	0.73	0.50	0.84
Inhibition	94	0.12	0.15	0.08	0.12	0.25*	0.37	0.07	0.58
Sustained Attention to Response Task									
Correct	87	205.91	17.63	198.06	21.65	0.65**	0.74	0.55	0.84
Commissions	87	9.85	5.93	11.26	5.60	0.58**	0.72	0.57	0.82
Mean RT	87	0.43	0.09	0.45	0.10	0.62**	0.76	0.63	0.84
RTCV	87	0.29	0.10	0.34	0.13	0.46**	0.58	0.33	0.73
Anticipations	87	6.25	11.59	10.95	13.88	0.64**	0.75	0.57	0.84
Omissions	87	2.99	4.12	4.72	6.69	0.43**	0.54	0.30	0.70
Updating Letter Memory Task									
Letters recalled	86	35.31	10.48	38.95	12.61	0.69**	0.79	0.65	0.87
Stop-Signal Task									
p(r|s)	112	52.02	15.21	52.52	14.22	0.69**	0.82	0.73	0.87
SSD	112	323.38	167.84	347.64	174.76	0.68**	0.81	0.72	0.87
SSRT	112	303.44	81.71	285.80	74.81	0.57**	0.71	0.58	0.80
No signal RT	112	627.55	133.31	634.58	141.21	0.64**	0.78	0.68	0.85

Note. RTCV = Response time coefficient of variability; p(r|s) = probability of responding on stop-signal trials; SSD = stop-signal delay; SSRT = stop-signal reaction time

*Correlation is significant at the 0.05 level

** Correlation is significant at the 0.001 level

Reliability scores from the inhibition measure are more stable across studies than those observed in the Stop-Signal Task and can be interpreted as moderate in most cases except for the one reported by Strauss et al. (2005), which presents a low correlation for a test-retest design. The reliability found in our study for this particular measure is poor, indicating an inadequate measure of inhibition.

Although the ICC scores of the congruent and incongruent trials of the Stroop task from the Ecuadorian sample are moderate to good, according to the 95% confidence interval range, the ICC for the inhibition score drops drastically, occupying most of the range considered as poor reliability (from 0 to 0.49). It is not surprising to find lower reliabilities for scores obtained indirectly as these tend to include the measurement error from the direct scores they are derived from (Caruso, 2004; Thomas & Zumbo, 2012). However, this is not sufficient to explain such low reliability in the Ecuadorian sample, mostly when the reliabilities found in the other studies are notably higher. One possible explanation for this is that the high number of participants evaluated in each session affected their performance in various ways. For example, it is possible that participants could have been distracted by the presence of other participants or could even have tried to "peek" at other screens, thus distracting them from their own task. This situation could affect participants' performance and the reliability of the tasks.

The reliability found in the SART, as measured by the ICC, is similar to those found in other reliability studies from different tasks. This means that the number of errors by commission is a reliable score in the Ecuadorian sample, or at least as reliable as scores from cognitive tasks tend to be. In the case of the other three scores proposed by Cheyne et al. (2009), namely the RTCV, number of anticipations, and number of omissions, only the number of anticipations provided a moderate to good reliability. The other two scores showed a poor to moderate reliability. It is possible that the group format of evaluation used in the present study also affected the performance of this task, especially when its monotonous nature can divert participants' attention to other participants or screens. Nevertheless, we recommend that if the RTCV and omissions scores are going to be used, efforts should be made to improve task reliability. One way to do this could be to increase the number of trials used on the task. One of the advantages of the SART is that, when following the original author's design (number of trials, duration of stimulus presentation, interstimulus interval, etc.), it usually takes around 5 minutes to be completed, thus, increasing this number would not be difficult.

Test-retest reliability information about the SART and the ULMT is more limited than the available for the Stroop and Stop-Signal task; in fact, we could only find one previous study reporting this information for the SART (Robertson et al., 1997), and none for the ULMT, making comparison between different populations impossible. Despite this limitation, the present study still represents an important contribution, as the ICC reliability of both tasks are presented here. The ULMT, in the format used by Miyake and Friedman (2012) in their studies about executive functions, has already been used by several laboratories (Dahlin et al., 2008; St Clair-Thompson & Gathercole, 2006) as it provides a straightforward assessment of the updating process of WM, however, to our knowledge the stability of its score across time has not been tested until now. According to our findings, this task can be used as a reliable measure of the updating capacity. In the case of the SART, the present study provides an ICC estimate which is a better reliability measure than the interclass correlation provided by the original author.

This study has several methodological limitations. Procedural differences between the pretest and retest sessions, with all tasks completed in a single session for the pretest but separately during the retest, may have introduced confounds such as fatigue, motivation changes, and variability in task order, potentially affecting performance and reliability estimates. These differences also raise concerns about assuming equivalent testing conditions when calculating ICCs, which rely on consistent measurement conditions. Additionally, technical issues during data collection led to some data loss, reducing the robustness and representativeness of the sample. Another limitation is the lack of convergent validity analysis, which would provide evidence on whether the tasks effectively measure the intended cognitive processes. While this study focused on test-retest reliability in a non-WEIRD population, future research should investigate the validity of these tasks to enhance their cross-cultural applicability and psychometric robustness.

In relation to our sample, it is important to note that, although the university students from the University of Guayaquil present important differences from those of "WEIRDer" countries, these are still young people who had the ability to finish secondary studies and who are interested in obtaining a university degree. This means that these students have to some extent overcome the difficulties present in the general Ecuadorian

population, namely, extreme poverty, violence in low-income neighbourhoods, and possible nutritional problems in low-income families. In this sense, they only represent the Ecuadorian population that have had access to higher education, making them share some commonalities with samples from studies using WEIRD populations. Nevertheless, this population shares cultural features that are distinct from the typical WEIRD population (Capella et al., 2019), providing a valuable source of information on the effects of such cultural variables on cognitive performance.

In general, our data suggest that the reliability estimates of the tasks with the Ecuadorian population are very similar to those observed in WEIRD countries, indicating that the measures are stable across populations with important differences. This supports the notion that findings from cognitive research can be generalised to different populations and suggests cross-cultural commonality in underlying cognitive processes. Moreover, the study provides some grounding for future research to be developed in Ecuador and other LA countries, a region of very limited scientific production, particularly in the cognitive psychology field.

References

Atkinson, L., Leung, E., Goldberg, S., Benoit, D., Poulton, L., Myhal, N., … Kerr, S. (2009). Attachment and selective attention: Disorganization and emotional Stroop reaction time. *Development and Psychopathology*, *21*(1), 99–126. https://doi.org/10.1017/S0954579409000078

Bekker, E. M., Overtoom, C. C., Kenemans, J. L., Kooij, J. J., De Noord, I., Buitelaar, J. K., & Verbaten, M. N. (2005). Stopping and changing in adults with ADHD. *Psychological Medicine*, *35*(6), 807–816. https://doi.org/10.1017/S0033291704003459

Berberian, A. A., Gadelha, A., Dias, N. M., Mecca, T. P., Bressan, R. A., & Lacerda, A. T. (2015). Investigation of cognition in schizophrenia: Psychometric properties of instruments for assessing working memory updating. *Jornal Brasileiro de Psiquiatria*, *64*(3), 238–246. https://doi.org/10.1590/0047-2085000000084

Berchtold, A. (2016). Test–retest: Agreement or reliability? *Methodological Innovations*, *9*, 205979911667287. https://doi.org/10.1177/2059799116672875

Blaskey, L. G., Harris, L. J., & Nigg, J. T. (2008). Are sensation seeking and emotion processing related to or distinct from cognitive control in children with ADHD? *Child Neuropsychology*, *14*(4), 353–371. https://doi.org/10.1080/09297040701660291

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences*, *8*(12), 539–546. https://doi.org/10.1016/j.tics.2004.10.003

Brands, A. M. A., Biessels, G. J., De Haan, E. H. F., Kappelle, L. J., & Kessels, R. P. C. (2005). Effects of type 1 diabetes on cognitive performance. *Diabetes Care*, *28*(3), 726–735.

Capella, M., & Andrade, F. (2017). Hacia una psicología ecuatoriana: Una argumentación intergeneracional sobre la importancia de la cultura y la glocalidad en la investigación. *Teoría y Crítica de La Psicología*, *9*, 173–195.

Capella, M., Jadhav, S., & Moncrieff, J. (2019). History, violence and collective memory: Implications for mental health in Ecuador. *Transcultural Psychiatry*, 1–20. https://doi.org/10.1177/1363461519834377

Caruso, J. C. (2004). A comparison of the reliabilities of four types of difference scores for five cognitive assessment batteries. *European Journal of Psychological Assessment*, *20*(3), 166–171. https://doi.org/10.1027/1015-5759.20.3.166

Cheyne, J. A., Solman, G. J. F., Carriere, J. S. A., & Smilek, D. (2009). Anatomy of an error: A bidirectional state model of task engagement/disengagement and attention-related errors. *Cognition*, *111*(1), 98–113. https://doi.org/10.1016/j.cognition.2008.12.009

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Dahlin, E., Neely, A. S., Larsson, A., Bäckman, L., & Nyberg, L. (2008). Transfer of learning after updating training mediated by the striatum. *Science*, *320*(5882), 1510–1512. https://doi.org/10.1126/science.1155466

Dalgleish, T. (1995). Performance on the emotional stroop task in groups of anxious, expert, and control subjects: A comparison of computer and card presentation formats. *Cognition and Emotion*, *9*(4),

341–362. https://doi.org/10.1080/02699939508408971

De Zeeuw, P., Aarnoudse-Moens, C., Bijlhout, J., König, C., Post Uiterweer, A., Papanikolau, A., … Oosterlaan, J. (2008). Inhibitory performance, response speed, intraindividual variability, and response accuracy in ADHD. *Journal of the American Academy of Child and Adolescent Psychiatry*, *47*(7), 808–816. https://doi.org/10.1097/CHI.0b013e318172eee9

Diamond, A. (2013). Executive functions. *The Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Dias, N. M., Gomes, C. M. A., Reppold, C. T., Bastos, A. C. M. F., Pires, E. U., Carreiro, L. R. R., & Seabra, A. G. (2015). Investigação da estrutura e composição das funções executivas: Análise de modelos teóricos. *Psicologia - Teoria e Prática*, *17*(2), 140–152. https://doi.org/10.15348/1980-6906/psicologia.v17n2p140-152

Duncan, G. J., & Brooks-Gunn, J. (2000). Family poverty, welfare reform, and child development. *Child Development*, *71*(1), 188–196.

Ebert, K. D., & Kohnert, K. (2011). Sustained attention in children with primary language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *54*(5), 1372–1384. https://doi.org/10.1044/1092-4388(2011/10-0231)

Ecker, U. K. H., Lewandowsky, S., Oberauer, K., & Chee, A. E. H. (2010). The components of working memory updating: An experimental decomposition and individual differences. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(1), 170–189. https://doi.org/10.1037/a0017891

Epp, A. M., Dobson, K. S., Dozois, D. J. A., & Frewen, P. A. (2012). A systematic meta-analysis of the Stroop task in depression. *Clinical Psychology Review*, *32*(4), 316–328. https://doi.org/10.1016/j.cpr.2012.02.005

Fierro Bósquez, M. J., Olabarrieta-Landa, L., Christ, B. R., Arjol, D., Perrin, P. B., Arango-Lasprilla, J. C., & Rivera, D. (2024). Normative data for executive function tests in an Ecuadorian Waranka minority population. *The Clinical Neuropsychologist*, 1-21.

Franzen, M. D., Tishelman, A. C., Sharp, B. H., & Friedman, A. G. (1987). An investigation of the test-retest reliability of the Stroop Color-Word Test across two intervals. *Archives of Clinical Neuropsychology*, *2*, 265–272.

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., Defries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological Science*, *17*(2), 172–179.

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology. General*, *137*(2), 201–225. https://doi.org/10.1037/0096-3445.137.2.201

Gaertner, L., Sedikides, C., Cai, H., & Brown, J. D. (2010). It's not WEIRD, it's WRONG: When Researchers Overlook uNderlying Genotypes, they will not detect universal processes. *Behavioral and Brain Sciences*, *33*(2–3), 93–94. https://doi.org/10.1017/S0140525X10000105

Garavan, H. (1998). Serial attention within working memory. *Memory and Cognition*, *26*(2), 263–276. https://doi.org/10.3758/BF03201138

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Instituto Nacional de Estadistica y Censos. (2018). *Encuesta nacional de empleo, desempleo y subempleo (ENEMDU)*. Retrieved from https://www.ecuadorencifras.gob.ec/documentos/web-inec/POBREZA/2018/Diciembre-2018/201812_Pobreza.pdf

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*(1), 47–70. https://doi.org/10.1037/0096-3445.132.1.47

Kessler, Y., & Oberauer, K. (2014). Working memory updating latency reflects the cost of switching between maintenance and updating modes of operation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(3), 738–754. https://doi.org/10.1037/a0035545

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kuntsi, J., Stevenson, J., Oosterlaan, J., & Sonuga-Barke, E. J. S. (2001). Test–retest reliability of a new delay aversion task and executive function measures. *British Journal of Developmental Psychology*, *19*(3), 339–348.

Leppink, J., & Pérez-Fuster, P. (2017). We need more replication research – A case for test-retest reliability. *Perspectives on Medical Education*, *6*(3), 158–164. https://doi.org/10.1007/s40037-017-0347-z

Logan, G. D. (1994). On the ability to inhibit thought and action: A users guide to the stop-signal paradigm. *Inhibitory Processes in Attention, Memory, and Language*. https://doi.org/10.1016/j.jsat.2006.09.008

Logan, G. D., Schachar, R. J., & Tannock, R. (1997). Impulsivity and inhibitory control. *Psychological Science*, *8*(1), 60–64.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163–203.

Masuda, T., & Nisbett, R. E. (2001). Attending holistically versus analytically: Comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology*, *81*(5), 922–934. https://doi.org/10.1037/0022-3514.81.5.922

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*(1), 28–50. https://doi.org/10.1177/1088868310366253

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46.

McLoyd, V. C. (1998). Socioeconomic disadvantage and child development. *American Psychologist*, *53*(2), 185–204.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14. https://doi.org/10.1177/0963721411429458

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

Morris, N., & Jones, D. M. (1990). Memory updating in working memory: The role of the central executive. *British Journal of Psychology*, *81*, 111–121.

Nisbett, R. E. (2003). *The geography of thought: Why we think the way we do*. New York: Free Press.

Oyserman, D., & Lee, S. W. S. (2008). Does culture influence what and how we think? Effects of priming individualism and collectivism. *Psychological Bulletin*, *134*(2), 311–342. https://doi.org/10.1037/0033-2909.134.2.311.supp

Pardo, J. V., Pardo, P. J., Janer, K. W., & Raichle, M. E. (1990). The anterior cingulate cortex mediates processing selection in the Stroop attentional conflict paradigm. *Proceedings of the National Academy of Sciences*, *87*, 256–259. https://doi.org/10.1080/00107518308210682

Peirce, J. W. (2008). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10. https://doi.org/10.3389/neuro.11.010.2008

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115

Rivera, D., Perrin, P. B., Stevens, L. F., Garza, M. T., Weil, C., Saracho, C. P., ... & Arango-Lasprilla, J. C. (2015). Stroop color-word interference test: normative data for the Latin American Spanish speaking adult population. *NeuroRehabilitation*, 37(4), 591-624.

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., & Yiend, J. (1997). "Oops!": Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia*, *35*(6), 747–758. https://doi.org/10.1016/S0028-3932(97)00015-8

Rodríguez Barreto, L. C., Pulido, N. del C., & Pineda Roa, C. A. (2016). Propiedades psicométricas del Stroop, test de colores y palabras en población colombiana no patológica. *Universitas Psychologica*, *15*(2), 255. https://doi.org/10.11144/Javeriana.upsy15-2.ppst

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Siegrist, M. (1995). Reliability of the stroop test with single-stimulus presentation. *Perceptual and Motor Skills*, *81*(3 Pt 2), 1295–1298.

Siegrist, M. (1997). Test-retest reliability of different versions of the stroop test. *Journal of Psychology: Interdisciplinary and Applied*, *131*(3), 299–306. https://doi.org/10.1080/00223989709603516

Soreni, N., Crosbie, J., Ickowicz, A., & Schachar, R. (2009). Stop Signal and Conners' Continuous Performance Tasks: Test-retest reliability of two inhibition measures in ADHD children. *Journal of Attention Disorders*, *12*(9), 137–143. https://doi.org/10.1177/1087054708326110

St Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *Quarterly Journal of Experimental Psychology*, *59*(4), 745–759. https://doi.org/10.1080/17470210500162854

Strauss, G. P., Allen, D. N., Jorgensen, M. L., & Cramer, S. L. (2005). Test-retest reliability of standard and emotional Stroop tasks: An investigation of color-word and picture-word versions. *Assessment*, *12*(3), 330–337. https://doi.org/10.1177/1073191105276375

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Thomas, D. R., & Zumbo, B. D. (2012). Difference scores from the point of view of reliability and repeated-measures ANOVA: In defense of difference scores for data analysis. *Educational and Psychological Measurement*, *72*(1), 37–43. https://doi.org/10.1177/0013164411409929

Toplak, M. E., Bucciarelli, S. M., Jain, U., & Tannock, R. (2009). Executive functions: Performance-based measures and the behavior rating inventory of executive function (BRIEF) in adolescents with attention deficit/hyperactivity disorder (ADHD). *Child Neuropsychology*, *15*(1), 53–72. https://doi.org/10.1080/09297040802070929

United Nations Development Programme. (2018). *Human Development Indices and Indicators: 2018 Statistical Update*. New York. Retrieved from http://hdr.undp.org/sites/default/files/2018_human_development_statistical_update.pdf

Universidad de Guayaquil. (2016). Población Estudiantil – Universidad de Guayaquil. Retrieved September 3, 2019, from http://www.ug.edu.ec/poblacion-estudiantil/

Van Den Heuvel, D. M. J., Ten Dam, V. H., De Craen, A. J. M., Admiraal-Behloul, F., Olofsen, H., Bollen, E. L. E. M., … Van Buchem, M. A. (2006). Increase in periventricular white matter hyperintensities parallels decline in mental processing speed in a non-demented elderly population. *Journal of Neurology, Neurosurgery and Psychiatry*, *77*(2), 149–153. https://doi.org/10.1136/jnnp.2005.070193

Verbruggen, F., Logan, G. D., & Stevens, M. A. (2008). STOP-IT: Windows executable software for the stop-signal paradigm. *Behavior Research Methods*, *40*(2), 479–483. https://doi.org/10.3758/BRM.40.2.479

Weafer, J., Baggott, M. J., & De Wit, H. (2013). Test-retest reliability of behavioral measures of impulsive choice, impulsive action, and inattention. *Experimental and Clinical Psychopharmacology*, *21*(6), 475–481. https://doi.org/10.1037/a0033659

Wöstmann, N. M., Aichert, D. S., Costa, A., Rubia, K., Möller, H. J., & Ettinger, U. (2013). Reliability and plasticity of response inhibition and interference control. *Brain and Cognition*, *81*(1), 82–94. https://doi.org/10.1016/j.bandc.2012.09.010.