Online Assessment of Neuropsychological Tests: Is it Reliable?

Master Thesis Neuropsychology
Utrecht University

Name: Michael Slagter
Student Number: 6856012
Date: January 11, 2022

Supervisor:
Teuni ten Brink
Second Assessor:
Anne Marieke Doornweerd

Making public: yes

Introduction

Cognitive functioning is vulnerable to the effect of ageing, neurological impairments, and psychiatric impairments (Gage et al., 1995). Cognitive functioning plays an important role in a person's personal and professional life. Hence, assessment of it is important to gain insight into a person's existing potentials or acquired shortcomings. Neuropsychological assessment, depending on the nature and quantity of testing, takes 1,5 to 4 hours, and additional time to score and interpret the tests. Neuropsychological assessment can be administered in person versus at the participant's home, with versus without (online) supervision of the assessor, and using paper-and-pencil tests versus digitalized tests. In Dutch healthcare, patients visit the psychologist for neuropsychological assessment, which is mainly done by means of standardized pen-and-paper tasks.

Testing in person with supervision is done to ensure that the patient can take the tests in a quiet and standardized environment, where he or she receives clear instructions. By ensuring these standardized conditions, a valid and reliable conclusion can be drawn about the patient's cognitive functioning (Laatsch, 2002). In addition, observations can be made, and the psychologist can respond to the patient's emotional state (Institute of Medicine, 2015). Aside to these benefits, testing in person with supervision has some downsides. One of them is that it's time consuming for the patient, as well as the psychologist or researcher. The patient has to travel to the hospital or testing facility, and a psychologist or researcher is needed to take the test.

Nowadays, assessment is mainly done by means of standardized pen-and paper tasks. In the early 80s, computerized testing rapidly became popular, as availability of computer technology increased (Ryabik & Olson, 1985). Ryabik and Olson (1985) stated that computerized testing may be vulnerable to quasi objectivity and that accuracy may not always be justified. In addition, early research showed some subtle changes in the response of participants to a computer compared to testing in person, such as higher levels of state anxiety (Space, 1981). Since the 1980s however, technology has made significant developments and the computer is now a common domestic object. In more recent years, several studies have successfully digitalized existing neuropsychological tests (Moore et al., 2019; Spreij et al., 2020). Unsurprisingly, computerized assessment increases the feasibility of well-controlled and reliable test administration, which has proven to provide the opportunity to collect additional test data including work strategy and RT's (Choudeshwari et al., 2013). Another advantage of digitalized tests is that they could, potentially, be administered online, either at a test location or the participant's home, without presence of a psychologist or researcher.

In the past decade, studies have investigated the validity, reliability, and acceptance of self-administered, online neuropsychological testing (Feenstra et al., 2017, Chaytor et al., 2020). For example, the study of Feenstra et al., (2017) showed overall positive results with respect to usability of a newly developed Amsterdam Cognition Task, which participants had to complete from home. All participants were able to complete the test battery, consisting of 6 subtests, without the help of a test leader. However, 5 out of the 6 subtests did not meet the benchmark for test-retest reliability. Chaytor et al. (2020) found acceptable overall construct validity for her Test My Brain (TMB) battery, consisting of 5 subtests, including a digit span, vocabulary, and matrix reasoning test. Further, the study found associations with self-reported everyday functioning (ecological validity) that were comparable to in-person assessments, and the tests appeared to be appropriate for use in research applications. Limitations of this study were a small sample size, no calculations of test-retest reliability, and modest convergent validity. Both studies are amongst the first to thoroughly study self-administered online assessments. Although showing great potential, current research about self-administered online assessments is still scarce and has shortcomings. For example, much is still unknown regarding test-retest reliability, and whether it differs if a researcher is present during the administration.

In the current study, these challenges are investigated. Healthy individuals conducted a series of paper-and-pencil neuropsychological tests that were converted to a computerized counterpart. Online neuropsychological testing has several advantages compared to paper-and-pencil tests. First, patients or study participants who are not able to travel to the test facility (e.g. due to being bed-bound) or for whom this is not preferred (e.g. such as during a pandemic), will have the possibility to take the test at home. Second, time and money can be saved, as test results are automatically and reliably scored. Third, researchers will have the opportunity to reach significantly more people from all over the world by just sending them a link to assess tests or questionnaires, instead of inviting people to their laboratories. It was not a goal of this study to prove that online assessment can in the future completely replace paper-and-pencil tests, but given the benefits of online assessment, it could be a meaningful *addition* if paper-and-pencil testing has significant shortcomings for specific patients or situations.

Potential limitations of online testing are that the norm scores of paper- and pencil tasks will not automatically be applicable to the online counterparts, because the stimuli and means of assessment differ between paper-and-pencil tasks and online tests. Also, for online assessment, the participant needs to be able to use a computer, which may be harder for older people, or patients whose cognitive functioning has declined (Iverson et al., 2007). Another drawback of self-administered computerized testing is that the neuropsychologist or experimenter normally doesn’t observe the patient during administration. This potentially leads to less knowledge about the patient’s cognitive functioning compared to in-person administration, because less verbal and non-verbal signals may be picked up. In addition, patients might fail to read the instructions, misunderstand the instructions, or cheat during testing. For instance, a participant may write down words or numbers that they need to remember by heart, or look up the answer to a question on the internet.

The aim of this study was to find out whether online testing is reliable. To this aim the following research questions were answered: 1) Do people correctly follow the test instructions for the different online tests; and is this affected by the online presence of a researcher? 2.)What is the test-retest reliability of online neuropsychological tests? 3) Do people perform better on online tests when a researcher is present via a video call than when no researcher is present? In the current study, a video call was started during the assessment in one of the two experiment groups. Naturally, a video call during a self-administered assessment may partially eliminate the advantage of online testing being time extensive for the psychologist. However, a video call provides the opportunity to pick up verbal and non-verbal signals and the opportunity to clarify test instructions where needed in return. In this study, the experimenter did not provide test instructions, but was merely present during the assessment.

This study expected to point out issues with online testing, which could be improved for future research and application in clinical practice. This study evaluates the reliability of online assessment. Results of this study may be an addition to existing literature about online testing and open the way for further research towards modernized and automated online neuropsychological assessment, complementing the advantages, and eliminating the disadvantages of in-person, paper-and-pencil assessment.

Methods

Participants

A pilot study including 5 participants was conducted first to remove technical problems, improve instructions, and to decide whether the 7 selected tests were suitable for online administration. Eventually, all 7 tests were kept in the test battery. A group of 33 healthy Dutch speaking adult participants (age groups 18-70) years conducted these tests. For some participants, data on specific tests was inconclusive or missing and was not analyzed. Participants were recruited via social media channels, via friends and relatives of the researcher, and via the student recruitment system of Utrecht University (SONA). Participants conducted two sessions with the same tests. They were instructed to carry out the second part of the assessment within 6-10 days after the first session, if possible around the same time of the day as the first part. The second assessment was carried out to check for test-retest reliability and possible learning effects. Both assessments took approximately 45-60 minutes. Participants received no reward, but two winners of gift vouchers worth 25 euros were drawn from the pool. Participants used their own computer or laptop, mouse, keyboard, and internet connection. All instructions were in Dutch.

Procedure

There were two (between-subjects) assessment conditions for the first test session: one in which there was no researcher present and the participants had to watch and listen to the instructions by him/herself, and one situation in which there was a researcher present in the complete first assessment via a video call. The researcher did not give instructions or clarifications about the tests in any way.

Participants were asked to make sure the environment was quiet and non-distracting. The assessment was started via a link. Participants' display sizes may vary, so the number of pixels per cm was estimated on the basis of a calibration test, called the 'credit card check', which was carried out before the actual assessment started. In this check, participants were asked to change the dimensions of a rectangle on their screen until it perfectly aligned the dimensions of a credit card, which they could hold against the computer screen (Li et al., 2020). By doing this, it could be secured that stimuli were presented in the same size for all participants. Chances are minimal that the size of a participant's credit card exactly corresponded to the measurement of the credit card provided on the screen. This means that they didn't complete this check if their data on this check shows the exact number of pixels that was set as default in the program (500 pixels). The results of the credit card check (i.e. how many participants changed the dimensions of their on-screen rectangle) were investigated to answer the question if people follow test instructions.

To ensure participants heard the auditory (spoken) instructions, they were asked in the auditory instructions only to fill out a specific word. Also, participants were provided with the sum '5+2=…' on their screen. In the auditory instructions, they were asked to fill out the number '3' as an answer to that sum. Both auditory-only instructions ensured that participants could hear auditory instructions and were also implemented to check if participants followed test instructions. Another way to check if participants followed instructions, was a question about cheating after each test. The expectation was that participants were honest about whether they had cheated during the tests. Participants were provided with practice rounds for most of the tests to ensure the instructions were clear before the tests started.

Education level of participants was assessed using seven categories of a Dutch classification system, according to Verhage (1964), 1 being the lowest (less than primary school) and 7 being the highest (academic degree) (Verhage, 1964). These levels were converted into three categories: low (Verhage 1–4), average (Verhage 5), and high (Verhage 6–7).

Instructions

Each task started with written instructions was accompanied by an audio fragment in which the instructions were read out loud. In some tests, additional instructions were provided when the participant made specific rule violations or a mistakes.

Neuropsychological Tasks

This study consists of 7 neuropsychological tests: Line Bisection Task, Corsi Block Tapping Task, Star Cancellation, Digit Span, Simon Task, Greyscales Task, and Tower of London.

Line Bisection

The Line Bisection Task requires the participant to divide lines, drawn on the computer screen, in two equal parts. The task measures hemispatial neglect (Bailey et al., 2004). Kinsella et al. (1995) found a significant test-retest correlation of .64, indicating a moderate reliability. The outcome measure in the current study was the total mean deviation from the point the participant marked as the center of the line, to the actual center of the line. Measurements were in cm and ranged from less than -0.6 to more than 0.6 cm. A positive deviation means the participant draws the middle point at the right side of the actual middle, a negative deviation means the participant draws the middle point at the left side of the actual middle. A deviation more than 0.6 cm (or less than -0.6 cm) indicates serious hemispatial neglect (Zeltzer et al., 2008).

Corsi Block Tapping Task

In the Corsi Block Tapping, the participant repeats a sequence of visual stimuli in the presented order, or in the inverse order (Corsi, 1972). The original Corsi Block-Tapping task uses nine cubes placed on a wooden board to assess visuospatial working memory in the original test. In the current study, the blocks were presented on the computer screen and lighted up one by one. The digital version of the Corsi Block Tapping Task shows a reliability of .77, according to a study of Siddi et al. (2020). The outcome measure for the Corsi Block Tapping Task Forward was the total score, calculated by multiplying the maximum achieved block span with the total number of correctly recalled sequences across the whole task (Nitz, 2021). The scores on the Corsi Block Tapping Task Forward ranged from 0 (no correct sequences) to 144 (all 16 sequences of 9 different block spans were correct). The outcome measure for the Corsi Block Tapping Task Backward was the total score, calculated by multiplying the maximum achieved block span with the total number of correctly recalled sequences across the whole task (Borchert, 2021). The scores on the Corsi Block Tapping Task Forward can range from 0 (no correct sequences) to 126 (all 14 sequences of 9 different block spans were correct).

Simon Task

The Simon Task is an instrument to measure attentional demands and inhibitory control (Cevada et al., 2019). In the current study, the participants were required to respond with letter keys on a keyboard to non-spatial features of a stimulus (Kubo-Kawai & Kawai, 2010). A blue or red stimulus was randomly presented at either the left or right on the screen, but the position of the stimulus was always irrelevant. The participant was instructed to react to a red stimulus with the letter A, and to a blue stimulus with the letter L. Responses are generally faster when the red stimulus appears on the left (congruent response) than when it appears on the right (incongruent response), i.e., the Simon effect (Simon, 2011). Cevada et al. (2019) found a good measurement of reliability (Cronbach's α = .88). The outcome measure in this study was the reaction time to congruent stimuli, which can range from 0 ms (the participant pressed a key at the same time that a stimulus appeared) to 650 ms (the time limit that a participant was allowed to take to respond, set by the program).

Digit Span

The Digit Span uses escalating sequences of numbers from 1 to 9, presented in a randomized fashion, to assess verbal working memory. The participant must repeat these sequence of numbers in the presented order, or in the inverse order (Wechsler, 2008). The reliability of the WAIS-IV Digit Span Forward and Backward is .81 and .82, respectively (Wechsler, 2008). Outcome measures in this study were the number of correctly repeated sequences, which can range from 0 (no correctly repeated sequence) to 16 (all repeated sequences were correct) for both the Digit Span Forward and the Digit Span Backward.

Star Cancellation

The Star Cancellation Task involves cancellation of small stars randomly placed on the screen additionally containing distractors (large stars and letters), measuring unilateral visual neglect (Bailey et al., 2004). The best known measure for test-retest reliability, according to Bailey et al. (2004), comes from a sample size of only 10 subjects who performed the paper and pencil version. A test-retest reliability of .99 was found in this study. The outcome measure in the current study was a measure called *intersection rate,* which is the total number of potential crossing search patterns divided by the number of actual crossing lines in the test (Woods & Mark, 2007), indicating the quality of the search pattern the participant used to find all stars. The *intersection rate* is not a measure of neglect, but instead, this measure of search disorganization may reflect disturbance of an executive control mechanism (Woods et al., 2004). The range of the outcome measure can vary between 0 (no intersections were made) and 1 (all possible intersections were made, indicating a disorganized search pattern).

Greyscales Task

In the Greyscales Task, the participant is instructed to judge two rectangle-formed stimuli, placed above each other, which show a left to right gradient from light to dark. The task is highly sensitive to unilateral hemispheric damage (Mattingley et al., 2004) The two stimuli are mirror-reversed in terms of gradient (greyscales). The instructions are to judge which one of the stimuli is darker on average. In their original study, Mattingly et al. (1994) found that healthy adults showed a bias to choose the stimulus that's dark on the left side, even though the stimuli are mirrored versions of each other. Märker et al. (2019) found a test-retest reliability of the Greyscales Task of .64 in healthy older adults. The outcome measurement was the bias score, the proportion of trials in which the stimulus that was dark on the left side was chosen. The scores can range from 0 to 1. A score between 0 and 0.5 means a bias towards the left side, a score between 0.5 and 1 means a bias towards the right side of the screen.

Tower of London

In the original Tower of London Task, participants are presented with three rods with disks of different color and size. Participants are instructed to place the disks on the rods in specific positions, according to an example. They have to follow two rules, namely that larger disks cannot be placed on a smaller disk and only one disk can be moved at the time. Participants have to use as few steps as possible. The task difficulty is increased by each trial (Marchegiani et al., 2010). In the current study, a picture of rods and disks was presented, and participants had to move the disks by dragging them with their computer mouse. Köstering et al. (2015) found adequate test-retest reliability of experimental utility of the Tower of London Task of between .27 and .74 on several outcome measures. The outcome measurement in the current study was the cumulated points for all solved problems. More points were given when the participant solved the problem in the optimal amount of steps. More points were given when the difficulty of the solved problems increased (e.g. 2 points were rewarded for solving problem 1-2 in the optimal amount of steps, whereas 4 points were rewarded for solving problem 5-9 in the optimal amount of steps), but taking more steps decreased the amount of rewarded points per problem (Borchert, 2020).

Analyses

Raw data from the Inquisit data files was preprocessed using a custom made MATLAB R2020b script (Mathworks, 2021). SPSS Statistics version 27 was used for statistical analyses (IBM, 2021). For all analyses, results were considered statistically significant at alpha < .05.

Demographic characteristics

Demographic characteristics were compared between groups (i.e. with or without researcher). Sex, level of education, and handedness were compared between groups (i.e. with and without researcher) by means of Chi-square tests. Age was compared between groups (i.e. with and without researcher) by means of an independent t-test.

Do people follow test instructions?

The number of participants who followed the experiment instructions was compared between the conditions with or without the researcher. The experiment instructions consisted of the question about the sum (correct or incorrect), the question about which word was played to the participant (correct or incorrect), and whether the credit card check was performed (yes or no). To compare the groups, a Chi-square test was executed per question. To check for cheating, the answer to the corresponding question was compared between groups with and without researcher. After each task, the participants were provided with an open text block in which they could write down their test strategy. This is done to provide the researcher with qualitative information regarding the method the participant used in completing the tests. The participants could formulate their strategy in an open text box rather than checking a box with a prior formulated strategy that comes closest to their own, because in an open text box, the participant could describe their strategy in more detail. No statistical analyses were carried out on this data, but answers were analyzed for explorative purposes.

Test-retest reliability of online neuropsychological tests

To analyze the test-retest reliability of online tests, the performance on the first and second test session were compared by calculating Spearman's rho. The most commonly used interpretation of the r coefficient comes from Dancey & Reidy (2008), according to whom a correlation between .10 and .40 is considered weak, between .40 and .70 is considered moderate, and from .70 and higher it is considered strong. In this study, a strong test-retest correlation was needed, which means the correlation needed to be at least .70 for each test separately. If the correlation coefficient was lower than .70, the test-retest reliability was considered insufficient for this study and the tests would require improvement. Correlation coefficients of assessments with and without researcher were not compared to obtain maximum power.

Presence of researcher

To analyze whether people performed better on online tests when a researcher is present via a video call than when no researcher is present, all outcome measurements were compared between the groups (i.e. with and without researcher; subjects factor) and between the two test sessions (i.e. session 1 and session 2; within subjects factor) by means of a repeated measures ANOVA. This study expected to find better performance in the group where a researcher was present, compared to the group where no researcher was present.

Results

Demographic characteristics

Table 1 lists the demographic characteristics. The groups (i.e. with and without researcher) did not differ from each other regarding sex, handedness, age, and level of education.

Table 1:

Demographic characteristics, split per group (i.e. with and without researcher)

	Group with researcher (percentage)	Group without researcher (percentage)	Statistical comparison between groups
Group size	16	17	
Sex, Men	6 (42.86%)	8 (57.14%)	$X^2(1) = 0.58$,
Sex, Women	10 (52.63%)	9 (47.37%)	$p = .308$
Handedness, left	2 (66.66%)	1 (33.33%)	$X^2(1) = 0.44$,
Handedness, right	14 (46.66%)	16 (53.33%)	$p = .509$
Age, mean	34.9	32.1	$t(31) = 0.68$, $p = .504$
Education Average (Verhage 5)	1 (25.00%)	3 (75.00%)	$X^2(1) = 1.42$, $p = .491$
Education High (Verhage 6–7)	15 (51.72%)	14 (48.28%)	

Do people follow test instructions?

In groups with and without researcher, the performance for the sum, sound, and credit card check were compared. Regarding the sum check, one participant in the first test session without researcher made two errors in answering the sum in accordance to the spoken instructions (i.e. 7 and 57), before giving the correct response. Regarding the sound check, one participant filled in 'geluid' (the Dutch word for sound), in the second session of the group with researcher. Within the credit card check, a total of 5 participants didn't change the dimensions of their on-screen credit card. These participants were divided over session 1 (2 participants: 1 in the group with researcher, and 1 in the group without researcher), and session 2 (3 participants: 1 participant in the group with researcher, and 2 in the group without researcher).

A total of 5 participants answered they had cheated during the tests. They were divided over session 1: 1 participant in the group with researcher during the Line Bisection Task, 1 participant in the group with researcher during the Digit Span Task, 2 participants in the group without researcher: 1 during the Corsi Block Task and 1 during the Digit Span Task, and 1 participant in session 2, in the group with researcher during the Corsi Block Task.

Originally, Chi-square tests would be carried out on data to check whether certain groups would significantly deviate from other groups in terms of following the instructions or cheating. However, the assumptions for the Chi-square tests were not met. Therefore, Fisher-exact tests were carried out on this data. The outcome measures of the sum check, the sound check, the credit card check, and the cheating check did not differ between the groups (i.e. with and without researcher).

Test-retest reliability of online neuropsychological tests

All test outcome measures were compared between the first and second assessment. There was a strong positive relation between session 1 and session 2 regarding the deviation at the Line Bisection task, r = .79, p < .001. This suggests a strong test-retest reliability. There were moderate positive relations between session 1 and session 2 regarding the Corsi Block Score Forward, r = .60, p < .001, the Corsi Block Score Backward, r = .57, p = .001, the Greyscales Bias, r = .57, p = .001, the Simon Task Reaction Time Congruent, r = .68, p = < .001, the Digit Span Score Forward, r = .47, p = .007, and the Digit Span Score Backward r = .42, p = .020. This suggests a moderately strong test-retest reliability. No significant correlation was found between session 1 and session 2 regarding Tower of London Task, r = .29, p = .114 and the Star Cancellation Intersection Rate , r = .04, p = .842.

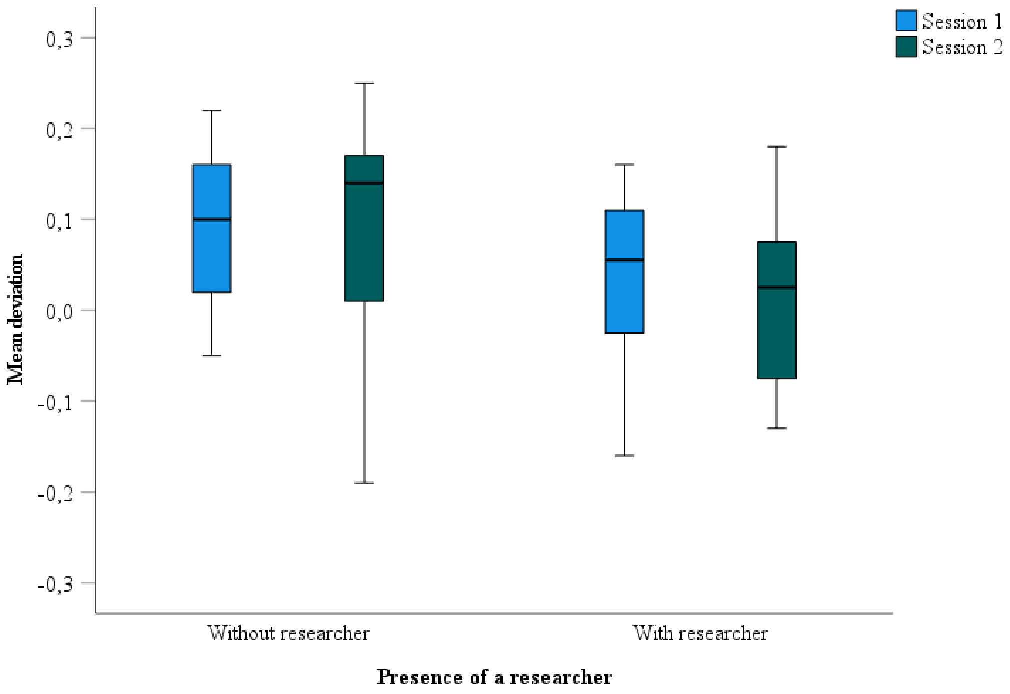
Presence of a researcher

Line Bisection Task

For the deviation on the Line Bisection task, there was no main effect of session, *F*(1,29) = 0.39, *p* = .536 (Figure 1), nor an interaction between session * presence of researcher, *F*(1,29) = 1.27, *p* = .269. There was a trend for an effect of presence of researcher, *F*(1,29) = 3.79, *p* = .062. The deviation was smaller when a researcher was present than when there was no researcher.

Figure 1:

Boxplot with scores on the Line Bisection Task



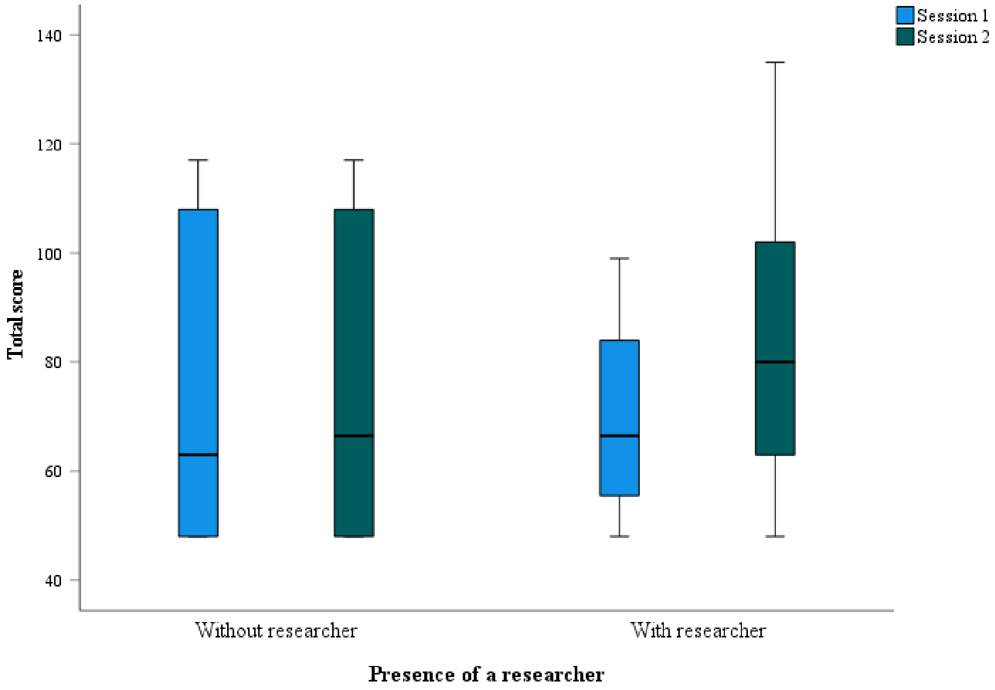
Note. This figure demonstrates the participants' score (i.e. the mean deviation to the middle of the line) on the Line Bisection Task in two separate sessions, displayed in boxplots. A score closer to 0 means the participant is better able to divide a line in two equals, compared to a high positive or negative score. A positive score means a deviation to the right, a negative score means a deviation to the left. The black horizontal line in each box represents the median of the participants' score in the corresponding group. The surrounding box represents the interquartile range. The minimum and maximum scores are represented by the upper and lower horizontal lines.

Corsi Block Tapping Task Forward

For the performance on the Corsi Block Tapping Task Forward Task, there was no main effect of session, *F*(1,29) = 2.36, *p* = .136 (Figure 2), no main effect of presence of researcher, *F*(1,29) = 0.04, *p* = .948, nor an interaction between session * presence of researcher, *F*(1,29) = 1.73, *p* = .199.

Figure 2:

Boxplot with scores on the Corsi Block Tapping Task Forward Task



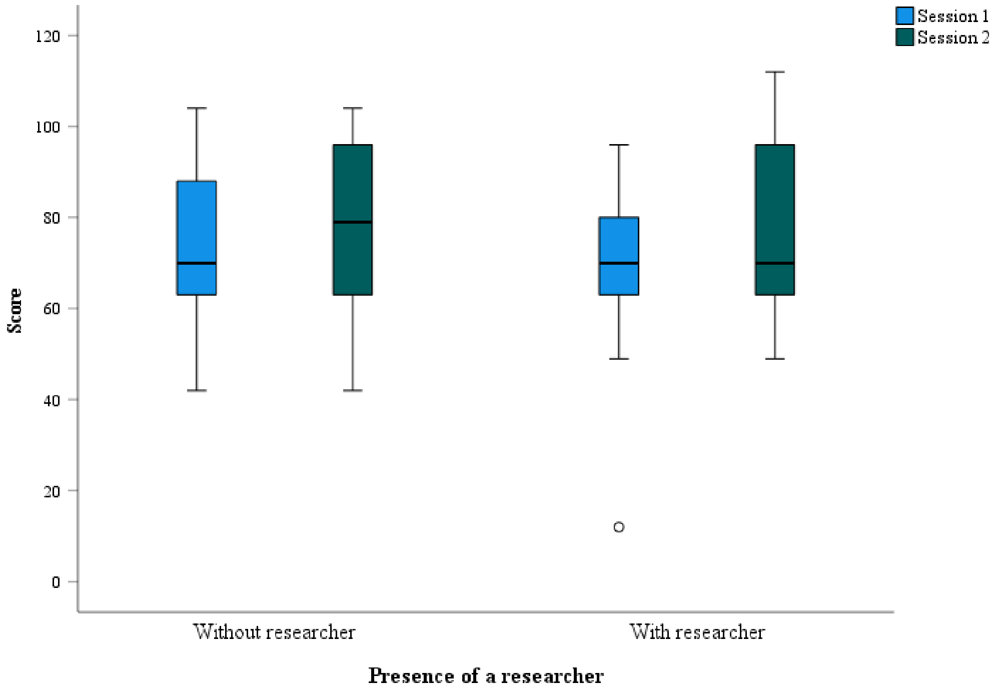
Note. This figure demonstrates the participants' total score (i.e. the longest remembered sequence of blocks, multiplied with the number of correctly remembered sequences) on the Corsi Block Tapping Task Forward in two separate sessions, displayed in boxplots. A high score means the participant has remembered more sequences of blocks. The black horizontal line in each box represents the median of the participants' score in the corresponding group. The surrounding box represents the interquartile range. The minimum and maximum scores are represented by the upper and lower horizontal lines.

Corsi Block Tapping Task Backward

For the performance on the Corsi Block Tapping Task Backward, there was no main effect of session, *F*(1,29) = 3.13, *p* = .088 (Figure 3), no main effect of presence of researcher, *F*(1,29) = 0.13, *p* = .719, nor an interaction between session * presence of researcher, *F*(1,29) = 0.11, *p* = .746.

Figure 3:

Boxplot with scores on the Corsi Block Tapping Task Backward Task



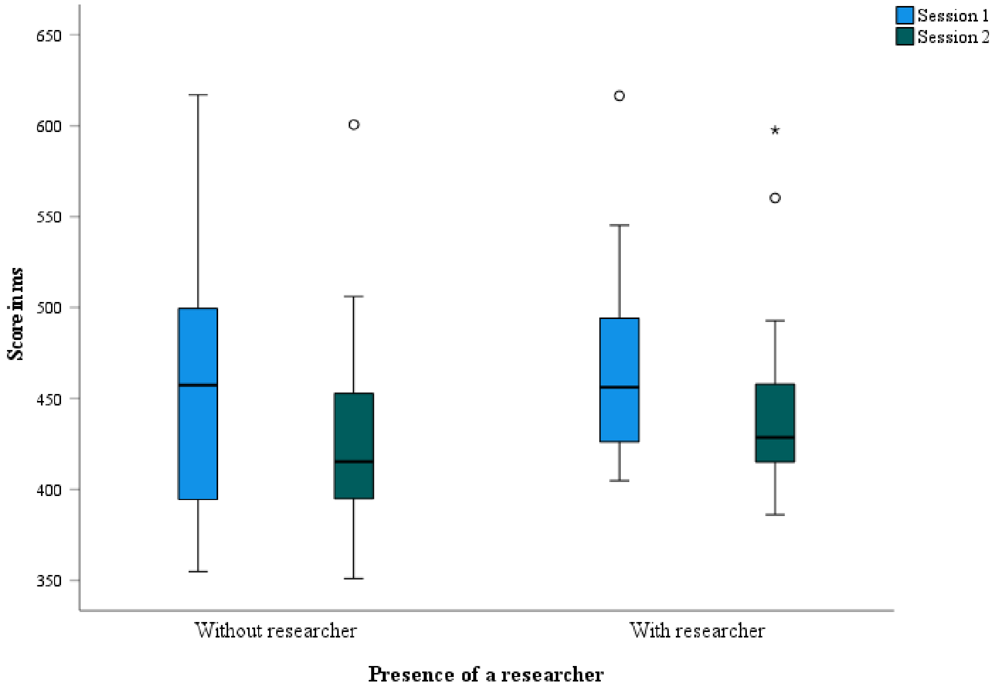
Note. This figure demonstrates the participants' total score (i.e. the longest remembered sequence of blocks, multiplied with the number of correctly remembered sequences) on the Corsi Block Tapping Task Backward in two separate sessions, displayed in boxplots. A high score means the participant has remembered more sequences of blocks. The black horizontal line in each box represents the median of the participants' score in the corresponding group. The surrounding box represents the interquartile range. The minimum and maximum scores are represented by the upper and lower horizontal lines. Non-significant outliers are displayed with dots.

Simon Task

For the Simon Task Congruent Reaction Time, there was a main effect of session, *F*(1,29) = 4.71, *p* = .039 (Figure 4), no main effect of presence of researcher, *F*(1,29) = 0.34, *p* = .564, and no interaction between session * presence of researcher, *F*(1,29) = 0.05, *p* = .831. Reaction times were faster during the second test administration.

Figure 4:

Boxplot with scores on the Simon Task Congruent Reaction Time



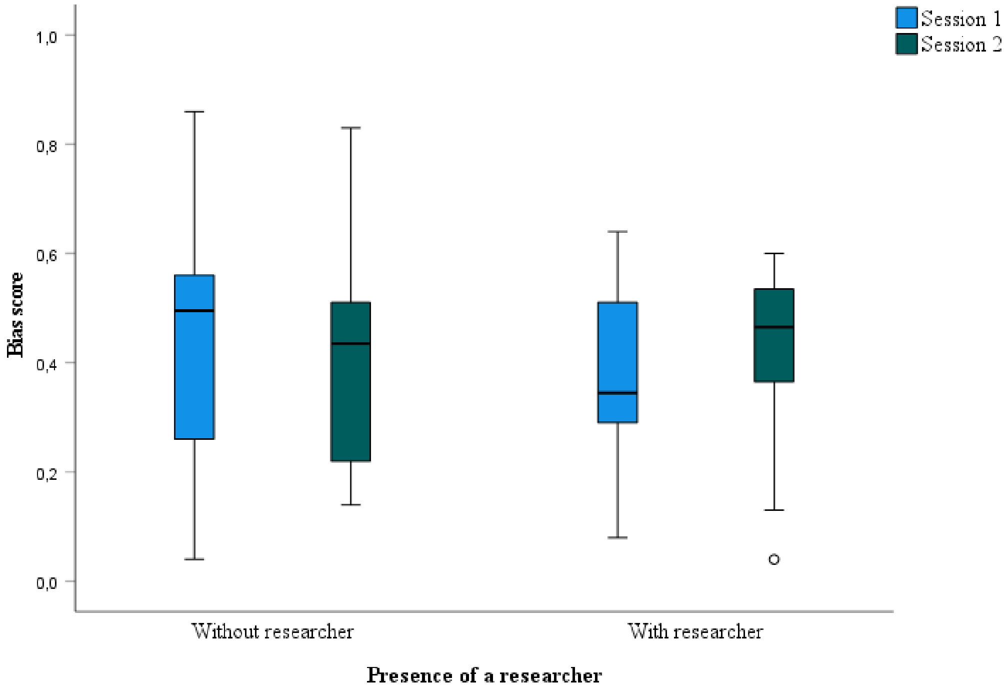
Note. This figure demonstrates the participants' score (i.e. the reaction time in ms) on the Simon Task Congruent in two separate sessions, displayed in boxplots. A lower score means the participant's reaction is quicker to a congruent stimulus (i.e. the stimulus appears on the same side of the screen as the corresponding finger is located). The black horizontal line in each box represents the median of the participants' score in the corresponding group. The surrounding box represents the interquartile range. The minimum and maximum scores are represented by the upper and lower horizontal lines. Non-significant outliers are displayed with dots, significantly deviant outliers are displayed with stars.

Greyscale Task

For the bias score on the Greyscales Task, there was no main effect of session, *F*(1,29) = 0.01, *p* = .947 (Figure 5), no main effect of presence of researcher, *F*(1,29) = 0.21, *p* = .649, nor an interaction between session * presence of researcher, *F*(1,29) = 2.30, *p* = .141.

Figure 5:

Boxplot with scores on the Greyscales Task



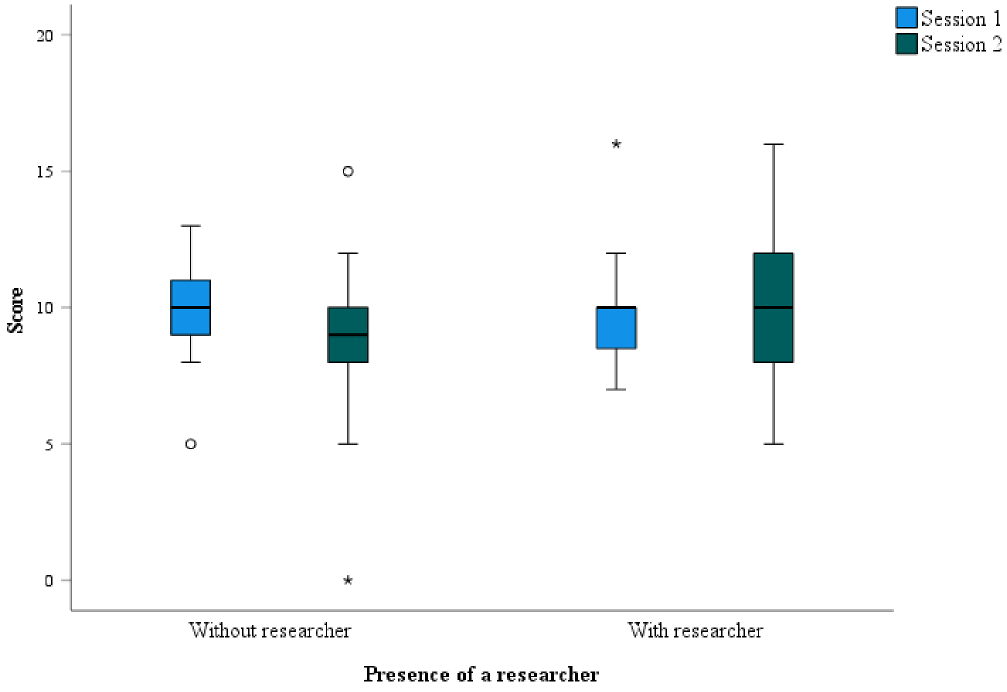
Note. This figure demonstrates the participants' bias score on the Greyscale Task in two separate sessions, displayed in boxplots. A score between 0 and 0.5 means a bias towards the stimulus that is dark on the left side, a score between 0.5 and 1 means a bias towards the stimulus that is dark on the right side of the screen. The black horizontal line in each box represents the median of the participants' score in the corresponding group. The surrounding box represents the interquartile range. The minimum and maximum scores are represented by the upper and lower horizontal lines. Non-significant outliers are displayed with dots.

Digit Span Task Forward

For the performance on the Digit Span Task Forward, there was no main effect of session, *F*(1,30) = 0.31, *p* = .581 (Figure 6), no main effect of presence of researcher, *F*(1,30) = 0.51, *p* = .481, nor an interaction between session * presence of researcher, *F*(1,30) = 1.29, *p* = .265.

Figure 6:

Boxplot with scores on the Digit Span Task Forward



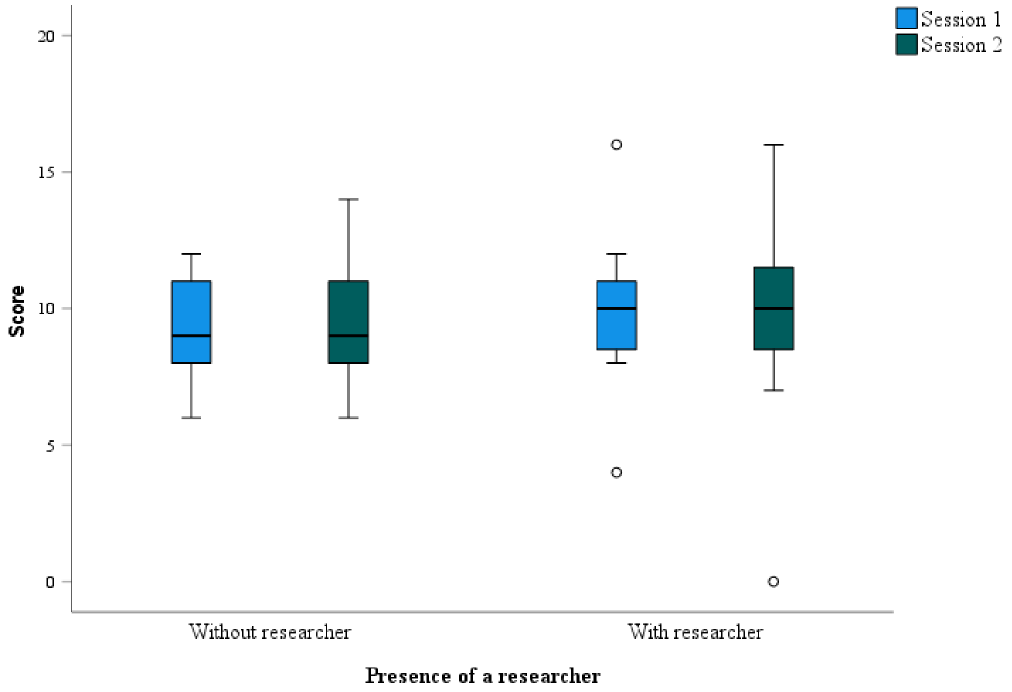
Note. This figure demonstrates the participants' score (i.e. the total number of correctly remembered digit sequences) on the Digit Span Task Forward in two separate sessions, displayed in boxplots. A higher score means a higher number of correctly remembered digit sequences. The black horizontal line in each box represents the median of the participants' score in the corresponding group. The surrounding box represents the interquartile range. The minimum and maximum scores are represented by the upper and lower horizontal lines. Non-significant outliers are displayed with dots, significantly deviant outliers are displayed with stars.

Digit Span Task Backward

For the performance on the Digit Span Task Backward Task, there was no main effect of session, *F*(1,30) = 0.22, *p* = .639 (Figure 7), no main effect of presence of researcher, *F*(1,30) = 0.28, *p* = .600, nor an interaction between session * presence of researcher, *F*(1,30) = 0.02, *p* = .895.

Figure 7:

Boxplot with scores on the Digit Span Task Backward Task



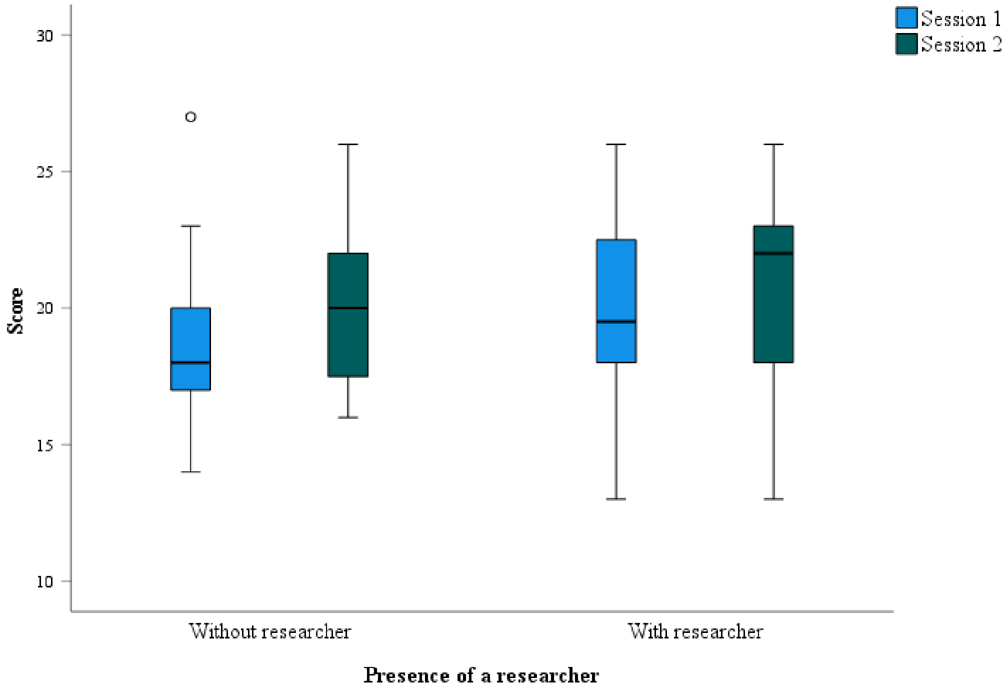
Note. This figure demonstrates the participants' score (i.e. the total number of correctly remembered digit sequences) on the Digit Span Task Backward in two separate sessions, displayed in boxplots. A higher score means a higher number of correctly remembered digit sequences. The black horizontal line in each box represents the median of the participants' score in the corresponding group. The surrounding box represents the interquartile range. The minimum and maximum scores are represented by the upper and lower horizontal lines. Non-significant outliers are displayed with dots.

Tower of London Task

For the performance on the Tower of London Task, there was no main effect of session, *F*(1,30) = 1.83, *p* = .187 (Figure 8), no main effect of presence of researcher, *F*(1,30) = 0.45, *p* = .508 nor an interaction between session * presence of researcher, *F*(1,30) = 0.16, *p* = .691.

Figure 8:

Boxplot with scores on the Tower of London Task



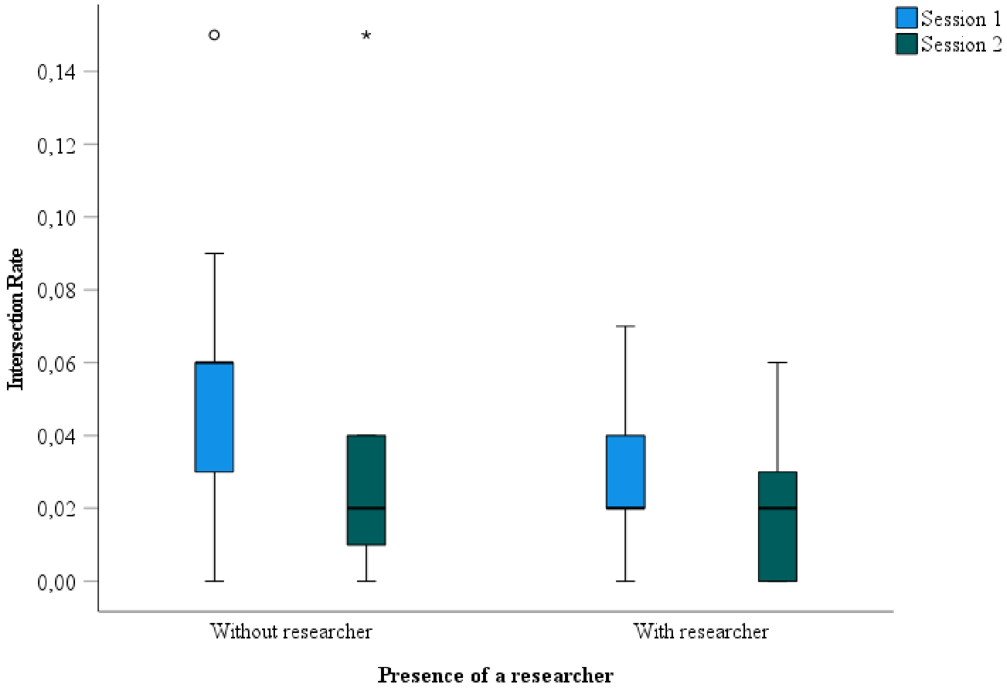
Note. This figure demonstrates the participants' score (i.e. the cumulated amount of points per specific solved problem) on the Tower of London Task in two separate sessions, displayed in boxplots. A higher score means a higher number of correctly solved problems. The black horizontal line in each box represents the median of the participants' score in the corresponding group. The surrounding box represents the interquartile range. The minimum and maximum scores are represented by the upper and lower horizontal lines. Non-significant outliers are displayed with dots.

Star Cancellation Task

For the performance on the Star Cancellation Task (i.e. the intersection rate), there was a significant effect of presence of researcher, $F(1,30) = 6.72$, $p = .015$ (Figure 9). The intersection rate was smaller when a researcher was present than when there was no researcher. There was a trend for main effect of session, $F(1,30) = 3,76$, $p = .062$, and there was no interaction between session * presence of researcher, $F(1,30) = 0.96$, $p = .336$.

Figure 9:

Boxplot with scores on the Star Cancellation Task



Note. This figure demonstrates the participants' intersection rate (i.e. the number of intersecting lines, deviated by the total number of potential intersecting lines) on the Star Cancellation Task in two separate sessions, displayed in boxplots. A lower score means a less intersections. The black horizontal line in each box represents the median of the participants' score in the corresponding group. The surrounding box represents the interquartile range. The minimum and maximum scores are represented by the upper and lower horizontal lines. Non-significant outliers are displayed with dots, significantly deviant outliers are displayed with stars.

Discussion

The main goal of this study was to find out if online testing is reliable. The test-retest reliability of online tests and the presence of a researcher are discussed separately.

A test-retest reliability of more than .70 is considered high (Dancey & Reidy, 2008) and was high enough for the purpose of the current study. In only one task (Line Bisection), the relation between the first en second administration was strong enough. The moderate test-retest reliability of one task (the Greyscale Task) in the current study corresponds to the test-retest reliability found in the study of Märker et al. (2019). In general, the test-retest reliability found in most of the neuropsychological tests in this study was lower than found in other studies (e.g. Siddi et al., 2020; Wechsler, 2008; Cevada et al., 2019; Bailey et al., 2004; Köstering et al., 2015). There are several potential contributing factors that can explain a lower test-retest reliability such as fatigue, low blood sugar, and specific mood and feelings in one of the sessions (Polit, 2014). Although these factors could contribute to the lower test-retest reliability found in the current study, it is unlikely that the group of participants in this study would display these factors more than in any other study.

In her paper, Polit (2014) found that rehearsal or practice effects can account for lower test-retest reliability. In the current study, the performance on the second test administration could be better because the participant may have remembered the tactics he/she used during the first administration. In a study of Rijnen et al. (2016), participants scored significantly higher on several neurocognitive domains at a retest of the same test battery 3 months after the initial assessment. The same result was found in a study of Hansen et al. (2016). A two week retest interval produced significant practice effects on all tests, particularly memory tests. In the current study, only one test found significant better performance on the second test administration. In addition, one test showed a trend for session. Although these results could be explained by practice effects, the other tests in the assessment did not significantly differ between the first and second administration. Therefore, practice effects could not explain the low test-retest reliability.

Another factor that could account for low test-retest reliability is environmental distraction. A study of Madero et al. (2021) shows that even during a 5-minute online eye-tracking test, 7.4% of the participants were distracted at least once. According to Madero et al. (2021), these rates of distraction frequency likely introduce enough uncertainty to preclude clinicians and researchers from using remote cognitive testing data from unsupervised tests without a reliable and validated form of quality assurance. Although participants in the current study were instructed to take the test in a quiet environment, it is likely that some participants could have been distracted during one or both test administrations. A study of Fox et al. (2009) shows that participants who were distracted by sending messages at the same time while performing a reading task took significantly longer to complete the task, indicating that distraction negatively affects efficiency. Accordingly, distractions in general during the test administration in the current study could have had a negative effect on maximum test performance. Most tests used in this online assessment are based on existing neuropsychological tests, but may measure more, or other constructs of neuropsychological functioning than their paper-and-pencil counterparts. As said, online tests are not necessarily comparable with paper-and-pen tasks, and comparisons between these must be made with caution.

For most tests, there was no effect for presence of a researcher. In two tasks (Line Bisection and Star Cancellation), there was a positive relation between presence of a researcher and performance on the task: in both tasks, the task performance was better with a researcher. Although this shows that performance can be improved when a researcher is present in some tasks, the most important finding is that that on most tests, performance was not affected by the presence of a researcher. These findings correspond to research of Lindman (2004), in which the presence of an observer or video camera proved to have no effect on participants who performed simple or complex tasks. Other research however, found a negative relation between the presence of a researcher and task performance on some, but not all tasks, either when the observer was a real person or a camera (Yantz & McCaffrey, 2007; Constantinou et al., 2005; Kehrer et al., 2000).

An important note is that most of the participants in the current study were young, highly-educated adolescents who were familiar with the use of a computer. Iverson et al. (2007) proved that people with frequent computer use performed better than people with some computer use on some tests requiring rapid visual scanning and keyboard use. Although the results of the current studies imply that the presence of an observer does not affect performance on most of the tests, the results do not necessarily apply to older or mentally declined people. Future research is needed to investigate whether the tests in the current study are suitable for clinical populations.

In general, participants followed the test instructions and completed the sum, credit card, and sound check. A few participants declared they had cheated during the tests. Although the amount of participants that didn't follow test instructions or cheated during tests was low, it can be advocated that these situations could be prevented in a traditional assessment with pencil-and-paper tests in a hospital. Many studies have investigated cheating by means of symptom validity tests (e.g. Greve et al., 2009; Beetar, 1995) in which subjects tend to underachieve or malinger symptoms, but studies investigating cheating to perform better on neuropsychological tasks are very scarce. To prevent cheating in future assessments, it may be necessary to clarify the test instructions regarding cheating even more. It could be valuable if participants or patients who cheat on purpose are detected in an early stage, for instance with a symptom validity test specially designed for cheating, or a diagnosis from a neuropsychologist. These patients could be deleted from the group of patients that qualify for an online assessment.

Limitations

This was an exploratory study with a relative small sample size of around 30 participants, potentially resulting in low statistical power. Nevertheless, this study can provide meaningful directions and knowledge for future research. An observation during the test administration was that some participants cheated without they declared this after the test. When asked about their actions, the participants declared that they did use the ‘forbidden’ technique, but that it was not cheating according to them. An important note is that most of the participants in the current study were young, highly-educated adolescents who were familiar with the use of a computer. Hence, the test results may not automatically be applicable for clinical populations. As noted before, there are a lot of points on which the current assessments needs to be improved before an online assessment can potentially be implemented in general clinical healthcare, for instance in terms of test-retest reliability, and better (video) instructions. Lastly, having no standardized, non-distracting environment could negatively impact test reliability. In the current study, the participants were given the instruction to take the test in a quiet, non-distracting environment only once. In future assessments, these instructions must be given more often and more emphatically.

Conclusions

The comparable test performance with or without a present researcher indicates that healthy participants don’t profit by a test leader in online neuropsychological assessment. Although this could mean that with clear instructions, online neuropsychological assessment could be carried out without a test leader and is therefore less time consuming, this statement must be made with care. In the current study, the researcher was merely present and did not help the patient or clarify the instructions. Further research is needed to investigate if participants take advantage of a researcher who is allowed to provide limited help or clarifications, as is done in traditional clinical practice. Further, some of the participants cheated or didn’t follow instructions, despite online presence of a researcher. This means that either test instructions must be made more clear, or patients who are prone to cheating must be removed from the patient group that qualifies for online assessment. In the current study, the test-retest reliability of the used tests was moderate and must be improved. After increasing the test reliability, further research is needed to clarify whether these tests are suitable for older, or mentally declined patients. In general, online assessment like the one that is carried out in the current study, is in an early stage of development. Many factors must be improved before implementation in clinical settings is justified. Perhaps the best way to look at online assessment at this moment is this: it should not yet be carried out to completely replace traditional assessment with pencil and paper tests on all patients, but could complement traditional assessment for specific patients. For instance, if a psychologist doubts if a younger patient shows temporary neurocognitive deficits as a result of a concussion, or if a highly educated patient shows questionable deterioration on some cognitive functions. In cases like these, online assessment could be a meaningful and possibly time saving first step to provide the neuropsychologist with information that could prove, or disprove his or her expectations.

Despite the aforementioned important shortcomings and imperfections at this stage, we’ve seen many upsides of online research that one day could contribute to, or even revolutionize neuropsychological assessment. With continued effort and refinements, online assessment shows to have great promise for the future.

References

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine, 18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Bailey, M. J., Riddoch, M. J., & Crome, P. (2004). Test–retest stability of three tests for unilateral visual neglect in patients with stroke: Star Cancellation, Line Bisection, and the Baking Tray Task. *Neuropsychological Rehabilitation*, *14*(4), 403–419. https://doi.org/10.1080/09602010343000282

Beetar, J. (1995). Malingering response styles on the memory assessment scales and symptom validity tests. *Archives of Clinical Neuropsychology, 10*(1), 57–72. https://doi.org/10.1016/0887-6177(94)e0005-a

Borchert, K. (2021, September 27). *User Manual for Inquisit's Corsi Block Tapping Task (Backwards).* Millisecond. Retrieved December 27, 2021, from https://www.millisecond.com/download/library/v6/corsiblocktappingtask/corsiblocktappingtask_backwards/corsiblocktappingtask_backwards/corsiblocktappingtask_backwards.manual

Borchert, K. (2020, January 14). *User Manual for Inquisit's Tower Test*. Millisecond. Retrieved December 21, 2021, from https://www.millisecond.com/download/library/v6/toweroflondon/towertest_d_kefs/towertest_d_kefs.manual

Chaytor, N. S., Barbosa-Leiker, C., Germine, L. T., Fonseca, L. M., McPherson, S. M., & Tuttle, K. R. (2020). Construct validity, ecological validity and acceptance of self-administered online neuropsychological assessment in adults. *The Clinical Neuropsychologist, 35*(1), 148–164. https://doi.org/10.1080/13854046.2020.1811893

Cevada, T., Conde, E., Marques, D., & Deslandes, A. C. (2019). Test-retest reliability of the simon task: a short version proposal. *Somatosensory & Motor Research, 36*(4), 275–282. https://doi.org/10.1080/08990220.2019.1689114

Choudeshwari, P. R., Kumar, A., Kumar, M., Harwani, Y., & Joshi, N. (2013). Comparative evaluation of computerized reaction times & critical flicker frequency with standard psychometric tests in diagnosis of minimal hepatic encephalopathy. *Journal of Clinical and Experimental Hepatology, 3*(1), S41–S42. https://doi.org/10.1016/j.jceh.2013.03.084

Constantinou, M., Ashendorf, L., & McCaffrey, R. J. (2005). Effects of a Third Party Observer During Neuropsychological Assessment. Journal of *Forensic Neuropsychology, 4*(2), 39–47. https://doi.org/10.1300/j151v04n02_04

Corsi, P. M. 1972. Human memory and the medial temporal region of the brain. *Dissertation Abstracts International, 34*(02), 891B. University Microfilms No. AAI05–77717

Dancey, C., & Reidy, J. (2008). *Statistics Without Maths For Psychology* (4th Revised edition). Pearson Education (Us).

Downloading IBM SPSS Statistics 27. (2021, May 21). IBM. Retrieved December 21, 2021, from https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-27

Feenstra, H. E. M., Murre, J. M. J., Vermeulen, I. E., Kieffer, J. M., & Schagen, S. B. (2017). Reliability and validity of a self-administered tool for online neuropsychological testing: The Amsterdam Cognition Scan. *Journal of Clinical and Experimental Neuropsychology, 40*(3), 253–273. https://doi.org/10.1080/13803395.2017.1339017

Fox, A. B., Rosen, J., & Crawford, M. (2009). Distractions, Distractions: Does Instant Messaging Affect College Students' Performance on a Concurrent Reading Comprehension Task? *CyberPsychology & Behavior, 12*(1), 51–53. https://doi.org/10.1089/cpb.2008.0107

Gage, R., Burns, J., Sellers, A. H., Roth, L., & Mittenberg, W. (1995). Approaches to memory assessment in the chronic psychiatric elderly. *Applied Neuropsychology*, *2*(3–4), 145–149. https://doi.org/10.1080/09084282.1995.9645352

Greve, K. W., Binder, L. M., & Bianchini, K. J. (2009). Rates of Below-Chance Performance in Forced-Choice Symptom Validity Tests. *The Clinical Neuropsychologist, 23*(3), 534–544. https://doi.org/10.1080/13854040802232690

Hansen, T., Lehn, H., Evensmoen, H., & Håberg, A. (2016). Initial assessment of reliability of a self-administered web-based neuropsychological test battery. *Computers in Human Behavior, 63*, 91–97. https://doi.org/10.1016/j.chb.2016.05.025

Institute of Medicine, Board on the Health of Select Populations, Testing, I. V. T. S. S. A. D. D., (2015). *Psychological Testing in the Service of Disability Determination*. Amsterdam University Press.

Iverson, G. L., Brooks, B. L., Ashton, V. L., Johnson, L. G., & Gualtieri, C. T. (2009). Does familiarity with computers affect computerized neuropsychological test performance?

Journal of Clinical and Experimental Neuropsychology, 31(5), 594–604. https://doi.org/10.1080/13803390802372125

Kehrer, C. A., Sanchez, P. N., Habif, U., Rosenbaum, G. J., & Townes, B. D. (2000). Effects of a Significant-Other Observer on Neuropsychological Test Performance. *The Clinical Neuropsychologist, 14*(1), 67–71. https://doi.org/10.1076/1385-4046(200002)14:1;1-8;ft067

Kinsella, G., Packer, S., Ng, K., Olver, J., & Stark, R. (1995). Continuing issues in the assessment of neglect. *Neuropsychological Rehabilitation, 5*(3), 239–258. https://doi.org/10.1080/09602019508401469

Köstering, L., Nitschke, K., Schumacher, F. K., Weiller, C., & Kaller, C. P. (2015). Test–retest reliability of the Tower of London Planning Task (TOL-F). *Psychological Assessment, 27*(3), 925–931. https://doi.org/10.1037/pas0000097

Laatsch, L. (2002). *Neuropsychological Assessment - an overview | ScienceDirect Topics*. Sciencedirect.Com. https://www.sciencedirect.com/topics/nursing-and-health-professions/neuropsychological-assessment

Li, Q., Joo, S. J., Yeatman, J. D., & Reinecke, K. (2020). Controlling for Participants' Viewing Distance in Large-Scale, Psychophysical Online Experiments Using a Virtual Chinrest. *Scientific Reports, 10*(1). https://doi.org/10.1038/s41598-019-57204-1

Lindman, Linda S. (2004). The effect of observational method and task complexity on neuropsychological test performance. *LSU Doctoral Dissertations*. 1248. https://digitalcommons.lsu.edu/gradschool_dissertations/1248

Madero, E., Anderson, J., Bott, N., Hall, A., Newton, D., Fuseya, N., Harrison, J., Myers, J., & Glenn, J. (2021). Environmental Distractions during Unsupervised Remote Digital Cognitive Assessment. *The Journal of Prevention of Alzheimer's Disease, 1–4.* https://doi.org/10.14283/jpad.2021.9

Marchegiani, A., Giannelli, M. V., & Odetti, P. R. (2010). The Tower of London test: A test for dementia. *Aging & Mental Health, 14*(2), 155–158. https://doi.org/10.1080/13607860903228804

Mark, V. W., Woods, A. J., Ball, K. K., Roth, D. L., & Mennemeier, M. (2004). Disorganized search on cancellation is not a consequence of neglect. *Neurology, 63*(1), 78–84. https://doi.org/10.1212/01.wnl.0000131947.08670.d4

Märker, G., Learmonth, G., Thut, G., & Harvey, M. (2019). Intra- and inter-task reliability of spatial attention measures in healthy older adults. *PLOS ONE, 14*(12), e0226424. https://doi.org/10.1371/journal.pone.0226424

Mattingley, J. B., Bradshaw, J. L., Nettleton, N. C., & Bradshaw, J. A. (1994). Can task specific perceptual bias be distinguished from unilateral neglect? *Neuropsychologia, 32*(7), 805–817. https://doi.org/10.1016/0028-3932(94)90019-1

Mattingley, J. B., Berberovic, N., Corben, L., Slavin, M. J., Nicholls, M. E. ., & Bradshaw, J. L. (2004). The greyscales task: a perceptual measure of attentional bias following unilateral hemispheric damage. *Neuropsychologia, 42*(3), 387–394. https://doi.org/10.1016/j.neuropsychologia.2003.07.007

Moore, R., Campbell, L., Delgadillo, J., Heaton, A., Leow, A., & Swendsen, J. (2019). S6. Digital Cognitive Assessment in Psychiatry Research. *Biological Psychiatry, 85*(10), S299. https://doi.org/10.1016/j.biopsych.2019.03.757

Neuropsychologisch onderzoek bij volwassenen. (n.d.). UMCG.Nl. Retrieved March 19, 2021, from https://www.umcg.nl/NL/Zorg/Volwassenen/zob2/Neuropsychologisch-onderzoek-bij-volwassenen/Paginas/default.aspx#:%7E:text=Het%20onderzoek%20wordt%20voor%20het,en%20eventuele%20adviezen%20te%20bespreken.

Nitz, D. (2021, September 27). *User Manual for Inquisit's Corsi Block Tapping Task.* Millisecond. Retrieved December 27, 2021, from https://www.millisecond.com/download/library/v6/corsiblocktappingtask/corsiblocktappingtask_forward/corsiblocktappingtask_forward/corsiblocktappingtask.manual

Polit, D. F. (2014). Getting serious about test–retest reliability: a critique of retest research and some recommendations. *Quality of Life Research, 23*(6), 1713–1720. https://doi.org/10.1007/s11136-014-0632-9

R2020b - Updates to the MATLAB and Simulink product families. (n.d.). Mathworks. Retrieved December 21, 2021, from https://nl.mathworks.com/products/new_products/release2020b.html

Rijnen, S. J. M., van der Linden, S. D., Emons, W. H. M., Sitskoorn, M. M., & Gehring, K. (2018). Test-retest reliability and practice effects of a computerized neuropsychological battery: A solution-oriented approach. *Psychological Assessment, 30*(12), 1652–1662. https://doi.org/10.1037/pas0000618

Ryabik, J. E., & Olson, K. R. (1985). Computerized testing. *Professional Psychology: Research and Practice, 16*(1), 6–7. https://doi.org/10.1037/0735-7028.16.1.6

Siddi, S., Preti, A., Lara, E., Brébion, G., Vila, R., Iglesias, M., Cuevas-Esteban, J., López-Carrilero, R., Butjosa, A., & Haro, J. M. (2020). Comparison of the touch-screen and traditional versions of the Corsi block-tapping test in patients with psychosis and healthy controls. *BMC Psychiatry, 20*(1). https://doi.org/10.1186/s12888-020-02716-8

Simon, J. R. (2011). “The Simon effect”: A potent behavioral mechanism. *Acta Psychologica, 136*(2), 181. https://doi.org/10.1016/j.actpsy.2010.04.007

Space, L. G. (1981). The computer as psychometrician. *Behavior Research Methods & Instrumentation*, *13*(4), 595–606. https://doi.org/10.3758/bf03202072

Spreij, L. A., Gosselt, I. K., Visser-Meily, J. M. A., & Nijboer, T. C. W. (2020). Digital neuropsychological assessment: Feasibility and applicability in patients with acquired brain injury. *Journal of Clinical and Experimental Neuropsychology*, *42*(8), 781–793. https://doi.org/10.1080/13803395.2020.1808595

Tijms, J. (2004). Verbal memory and phonological processing in dyslexia. *Journal of Research in Reading*, *27*(3), 300–310. https://doi.org/10.1111/j.1467-9817.2004.00233.x

Verhage, F. (1964). *Intelligence and age*. Assen, NL: van Gorcum [in Dutch]

Wechsler, D. A. (2008). *Wechsler Adult Intelligence Scale* (4th ed.). SanAntonio, TX: Psychological Corporation.

Woods, A. J., & Mark, V. W. (2007). Convergent validity of executive organization measures on cancellation. *Journal of Clinical and Experimental Neuropsychology, 29*(7), 719–723. https://doi.org/10.1080/13825580600954264

Yantz, C. J., & McCaffrey, R. J. (2007). Social Facilitation Effect of Examiner Attention or Inattention to Computer-Administered Neuropsychological Tests: First Sign that the Examiner May Affect Results. *The Clinical Neuropsychologist*, *21*(4), 663–671. https://doi.org/10.1080/13854040600788158

Zeltzer, L., Menon, A., Korner-Bitensky, N., & Sitcoff, E. (2008). *Line Bisection Test*. Strokengine. Retrieved December 19, 2021, from https://strokengine.ca/en/assessments/line-bisection-test/