cogent psychology

CLINICAL PSYCHOLOGY & NEUROPSYCHOLOGY | RESEARCH ARTICLE

Test administrator effects on cognitive performance in a longitudinal study of ageing

Marieclaire Overton[1]*, Mats Pihlsgård[1] and Sölve Elmståhl[1]

Received: 25 August 2016
Accepted: 07 November 2016
First Published: 12 November 2016

*Corresponding author: Marieclaire Overton, Division of Geriatric Medicine, Lunds University, Skånes University Hospital, Jan Waldenströms gata 35, CRC Building 28, fl.13, SE-205 02 Malmö, Sweden
E-mail: marie_claire.overton@med.lu.se

Reviewing editor:
Peter Walla, University of Newcastle, Australia

Additional information is available at the end of the article

Abstract: In longitudinal studies, changes in participants' cognitive performances can partly be attributed to variations in the testing situation (testing specific factors), such as, different test administrators or different testing environments. Focusing on test administrator influence, testing specific factors were examined in a Swedish longitudinal study of ageing. Test scores from 6,686 examinations revealed significant test administrator effects on all instruments measuring speed of processing, episodic memory and spatial ability. 1.4–3.5% of the total variation in test scores was explained by the factor attributed to the test administrator. Further, results indicated task familiarity on speed scores, and fatigue attributable to the time of day of the testing session for memory and spatial ability. Participants tested in the testing centre performed significantly better than participants who received home visits. This study provides evidence that testing specific factors are tangible concerns in longitudinal ageing studies.

Subjects: Neuropsychology; Cognitive Psychology; Cognitive Neuropsychology; Behavioral Psychology; Developmental Psychology

Keywords: test administrator effects; ageing; instrumentation; cognitive performance; longitudinal research

1. Introduction

Ageing studies that examine cognition, have a tradition of using longitudinal data collecting because of the advantages of this design, of which the most important is that it directly measures within

ABOUT THE AUTHOR

Marieclaire Overton holds a MSc in psychology from Lund University. She is currently doing her PhD in ageing and cognition at the Division of Geriatric Medicine at Lund university. Her research is based on data from the longitudinal study of ageing "Good ageing in Skåne" (GÅS). She has previously worked with testing GÅS participants' cognitive abilities. Whilst working she became interested in investigating test administrators influence participants' test performance. Cognitive researchers use participants' test results to distinguish cognitively impaired groups of participants from cognitively healthy groups. It is thus important to investigate whether the test results are reliable and free from exterior influences.

PUBLIC INTEREST STATEMENT

Today, it is customary to use psychological testing to measure elders' mental health. Potentially, test results assess level of mental health, and whether they should receive mental health care. Our research suggests that certain exterior aspects in the testing situation may influence test results, causing an unfair assessment. For instance, the test administrator can unconsciously assist or in severe cases hinder the senior being tested. Also, the time of day and which type of environment the elder is tested in may also be influential. Luckily, these influences were found to be small. Even so, researchers and psychologist should be aware of possible sources of influences, so that routines that reduce exterior influences are applied. Routines may include the test administrator following a strict and standardised testing protocol. By applying routines appropriately, we can be fairly confident that seniors will face a fair assessment.

person cognitive change. Despite the advantages, this design involves methodological choices made by the researcher and other practical issues that can lead to implications when interpreting results (Ferrer & Ghisletta, 2011). Methodological choices include procedures such as having several versions of the same test and offering home visits to participants. Further, longitudinal studies consisting of several follow-up periods and large study samples may use more than one test administrator to collect cognitive data. (Guo, 2013; Hofer & Schaie, 2001). These studies usually extend over a large period of time, and therefore personnel are often replaced throughout the course of the study. Longitudinal investigators have then a two-folded possible source of change essential to measure. Firstly, the change in the examined factor (e.g. cognitive function), and secondly the change of the testing setting (e.g. a different test administrator) for the test occasions (Guo, 2013). Inaccurate interpretation of cognitive maturation tendencies could be a consequence of not considering the latter types of influences (Hofer & Schaie, 2001), e.g. inaccurately interpreting participant's cognitive deterioration at the subsequent testing occasion. Thus, change in the testing setting is a potential concern in longitudinal studies of ageing and its influence is fundamental to measure (Hofer & Schaie, 2001; Schmidt & Teti, 2005; Shadish, Cook, & Campbell, 2002). This paper focuses on investigating potential sources of influences on participants' cognitive performances, that are tied to testing procedures. In particular, it will discuss influences attributable to the practice of using multiple test administrators.

1.1. Test administrator influence

Level of cognition is imperative to measure, and standardised tests as well as standardised procedures are often implemented to collect cognitive data. Testing procedures contain a component of interaction between the test administrator and the participant. Previous research has verified that test administrators have an influence on participants' cognitive performances (test administrator influence). This influence can be partially avoided by using strict and standardised procedures of test administration (Rousson, Gasser, & Seifert, 2002; Sattler & Theye, 1967). However, despite standardised testing procedures, some underlying influence may still remain. Sattler and Theye (1967) describe three possible sources of test administrator influence. These are: *departures from standardised procedure*, e.g. adapting testing instructions, *situational variables*, e.g. using praise and discouragement and the *experimenter variable*, involving attributes of the test administrator. Ethnicity (Marx & Goff, 2005; Samuel, 1977), sex (Ortner & Vormittag, 2011; Samuel, 1977), prior experience of administrating (Hoyt & Kerns, 1999; Lim, 2011), expectancies on participant's behaviour (Rosenthal & Fode, 1963), personal reactions to participant attributions, e.g. the halo effect (Hoyt, 2000; Thorndike, 1920), and attitude (Bookout & Hosford, 1969) have all been suggested to influence participant's performance.

Hoyt and Kerns (1999) suggest that *rater bias* (a form of test administrator influence) is ascribable to the characteristics of the testing instrument in combination with the experimenter variable. In their meta-analysis, 47% of variance in rating scores was found attributable to a combination of the experience of the rater and the inferential level of the instrument's rating system. The largest variance of rater bias was established for inexperienced raters using highly inferential rating systems. However, there was still a substantial amount of variance (19%) in ratings although the ratings were conducted by experienced raters. Despite the convincing evidence of test administrator influences on cognitive performance, general discussions on the subject in longitudinal studies and in gerontological research have been somewhat neglected (Guo & Bollen, 2013; Ortner & Vormittag, 2011). Longitudinal studies often use multiple test administrators to collect data, and systematic differences in testing procedures between these test administrators may exist. Hence, it is important to examine whether and to what extent test administrator influence exists in these types of studies.

1.2. Testing specific factors

Procedural choices and practicalities can influence participant's cognitive performance and procedures may vary within participant but also vary between participants. Procedural choices and practicalities may include changes of the test administrator, cognitive test batteries, the time of day the person is tested and in which type of testing environment the person was tested.[1] All of these factors are tied to the specific testing situation and are therefore referred to as *testing specific factors*. Testing specific factors, reasons for implementing these and suggestions on how they influence performance

Table 1. Testing specific factors in the GÅS-study, implementation reason and suggested latent factors

Testing specific factors	Implementation reason	Latent factors
Multiple test administrators	Practical: Change in personnel throughout the study	Test administrator influence, e.g. experience of test administrator
Different test versions	Reduce practice effects on subsequent testing occasions	Different difficulty levels of the test versions
Different orders of test battery administration	To reduce potential systematic fatigue effects	Fatigue due to serial testing or alertness in the beginning of the testing session.
Time of day of testing	Practical: To test many participants in the same day	Fatigue due to differences in circadian rhythms.
Testing setting	To boost participant rate/minimise selection bias	Discomfort/comfort of the participant

are discussed below and displayed in Table 1. These types of procedures are often implemented to reduce problems related to large population samples and to longitudinal data collecting. For example, a different version of the same test at the subsequent testing occasion can be used to reduce practice effects (Salthouse, 2014). Additionally, changing orders in which the tests are presented in a test battery can be used to reduce potential systematic fatigue effects on participants' performances due to serial testing (Laukka et al., 2013). There might also be practical motivations behind implementation of specific procedures, such as testing participants at different times of day or the need to use multiple test administrators to collect data. Moreover, home visits may be offered to participants unable to come to the test centre in order to avoid losing participants that are too frail to come to the testing centre, thus boosting participant rates and reducing selection bias (Lissner et al., 2003).

Testing specific factors do not intrinsically influence participant's performance, but indirectly affect test scores, due to latent factors. Latent factors are the suggested explanations behind how testing specific factors cause alteration in performance. Previous research on the latent factors, related to the testing specific factors, is scarce. It has however been observed that variations in test scores due to different versions of the same tests were explained by differences in difficulty levels between test versions. Also, variation of the order of which a test is presented in the test battery is designed to measure and reduce the effects of fatigue due to serial testing (Laukka et al., 2013). Hence, test score differences between test orders can indicate fatigue or alertness of the participant. Past research has found that variations in test performances can be seen due to the time of day of the testing, where age of the participant particularly matters (Blatter & Cajochen, 2007; May, Hasher, & Foong, 2005; Schmidt, Collette, Cajochen, & Peigneux, 2007). These differences are described as fatigue or alertness of the participants due to individual differences in circadian rhythms. Home visits in comparison to testing in the clinic have been found to positively influence test scores on the Mini mental state examination (MMSE). This is explained by the participant being more comfortable when tested at home (Shievitz, Tudiver, Araujo, Sanghe, & Boyle, 1998).

1.2.1. The present study

The overall aim of this study was to inspect to what extent the use of multiple test administrators together with other testing specific factors could influence participant's cognitive test scores. This was done in a longitudinal study of ageing, that used multiple test administrators, implemented several testing specific procedural choices and used tests designed to measure multiple cognitive domains; speed of processing, episodic memory and spatial ability. The first specific aim was to examine the extent of potential test administrator influence on participant's test scores. This was done by examining if any of the explained variance in test scores was attributed to the practice of using multiple test administrators. The second aim was to examine potential influences on test scores of all testing specific factors included in the study, such as different versions of the same test or time of day of testing (all displayed in Table 1).

2. Method

2.1. Sample

Participants were recruited through the Good Ageing in Skåne study (GÅS: Gott åldrande i Skåne), which is an ongoing longitudinal study in the southern part of Sweden, covering five municipalities including both rural and urban areas. Participants were randomly invited from the general population using the National Population Registry (Ekström & Elmståhl, 2006). Participants were examined by a physician, registered nurse and psychological test administrator. Dementia was diagnosed by the examining physician according to the Diagnostic and Statistical Manual of Mental Disorders-IV criteria, and medical records. Two questionnaires about health and lifestyle were provided. Full examination took 7 h and was conducted at one of the study's test centres. In order to avoid losing participants, participants unable to come to the test centre were offered an examination in their home. So far, four data collection sessions are complete and the fifth is ongoing. Re-examination was performed every third year for participants above 78 years of age and every sixth year for participants below 78 years of age. New participants were recruited every sixth year. Additional descriptions of the study design have been provided elsewhere (Ekström & Elmståhl, 2006; Lagergren et al., 2004).

90% (8,204) of 9,107 participant observations provided data from at least one cognitive test in the test battery, see Figure 1. The 8,204 participant observations in this sample were acquired from 4,897 unique participants. The sample contains participants observations from participants examined up to five times (see Table 2). Reasons for not taking part in the psychological testing session included dementia, over medication, illness, participation merely via telephone interview, and refusal. Further exclusion was based on well-known factors that affect cognitive performance. Participants with missing scores or scores <24 on mini mental state examination (MMSE; Folstein, Folstein, & McHugh, 1975), moderately and severely depressed participants, participants that had very poor eyesight, and non-native speaking participants with language difficulties were excluded. The evaluations whether poor eyesight and language difficulties affected the execution of the test were made by the test administrator. Furthermore, the incomprehension of the instructions of a specific test, as judged by a test administrator, led to exclusion for each test individually, in total 31 participant observations were therefore excluded. The excluded group was older (*M* = 80.60, *SD* = 10.63) than the main sample of participants (*M* = 71.34, *SD* = 10.12), and more likely to have lower education (mean years of education = 9.13, *SD* = 3.41), compared to the main sample

Figure 1. Flow chart of the sample with the exclusion criteria.

Note: MMSE = Mini mental state examination.

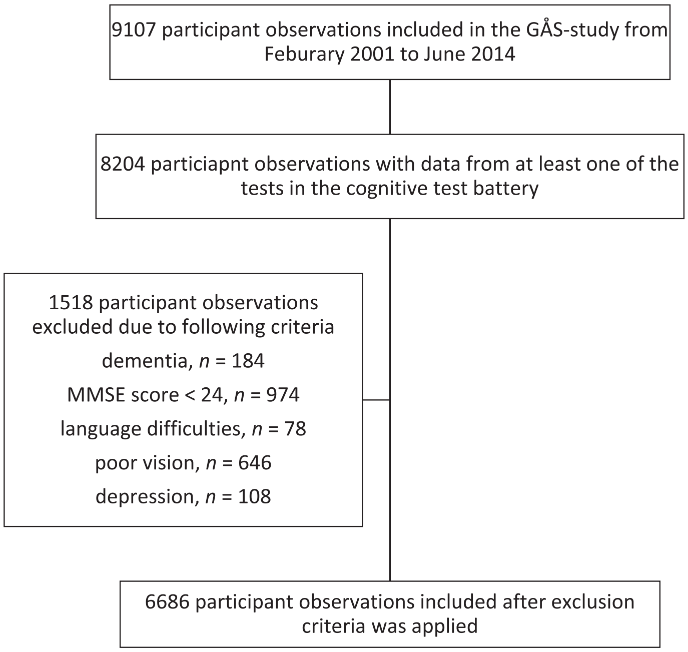
9107 participant observations included in the GÅS-study from Feburary 2001 to June 2014
8204 particiapnt observations with data from at least one of the tests in the cognitive test battery
1518 participant observations excluded due to following criteria
dementia, *n* = 184
MMSE score < 24, *n* = 974
language difficulties, *n* = 78
poor vision, *n* = 646
depression, *n* = 108
6686 participant observations included after exclusion criteria was applied

Table 2. Number of examination times for participants

	Number of participants	Percentage
Examination times		
1	2,611	53
2	1,523	31.1
3	542	11.1
4	184	3.8
5	37	.7
Total	4,897	100

Table 3. Sample characteristics after exclusion and descriptive statistics of participant observations and testing specific factors

	Mean	Standard deviation	Number of observations	Percentage
Age (years)	71.34	10.1		
Education (years)	10.90	3.79		
MMSE	27.62	1.69		
Sex				
Women			3,602	53.9
Men			3,084	46.1
Previous experience				
No experience at the time of examination			4,090	61.2
Previous experience			2,596	38.8
Time of day				
Morning			2,243	33.5
Prior lunch			2,820	42.2
Afternoon			1,618	24.2
Test order				
1			3,390	50.7
2			3,296	49.3
Test version				
1			2,248	33.6
2			2,219	33.2
3			2,219	33.2
Testing setting				
Test centre			5,921	88.6
Home visit			492	7.4

Notes: The 6,686 observations presented here were acquired from 4,090 distinct participants. MMSE = mini mental state examination.

(*M* = 10.90, *SD* = 3.79). There were more females in the excluded group (61% vs. 54%). After exclusion, 6,686 participant observations were included in the main sample. The following nine age groups were used; 60, 66, 72, 78, 81, 84, 87, 90, 93+. See main sample characteristics in Table 3. The study was approved by the regional ethics committee of Lund University (2002; registration No. LU 744–00) and all participants have provided written consent.

2.2. Test administrators

By June 2015, 21 (2 men, 19 women) test administrators have been involved in the study. Mean age was 32.8, median age was 30, age ranges from 25 to 50 years. The average duration of employment was 2.19 years, ranging between 0 and 10 years. Minimum requirement for serving as a test administrator was a bachelor's degree in behavioural sciences. The test administrators have tested between 42 and 1,503 participants.

Each administrator was provided with a detailed instruction handbook and had a minimum of two weeks training together with an experienced test administrator. The training weeks ended with a couple of practice sessions with a mock participant, before working with participants. The test administrator was observed by an experienced test administrator at the practice sessions, and during a real testing session. In addition, frequent meetings between test administrators and an experienced neuropsychologist took place every month, and more frequently if requested by the test administrator. This was to ensure standardised administration and consistency in test scoring. These meetings consisted of discussions of testing methodology and discussions of ambiguous responses from participants. To minimise expectancy effects, no information, prior to the testing session, on how returning participants performed on their last testing session was provided.

2.3. Cognitive assessment

The cognitive testing session was conducted by a test administrator and consisted of a standardised test battery, containing 12 neuropsychological tests, adapted for an older population. The examination time was about 1.5 h. Participants were either tested in the morning, between 8 and 10 am, or prior lunch, between 10 am and noon or in the afternoon, between 1 and 3 pm. To minimise systematic effects of fatigue due to serial testing, the test was administrated in two different orders in the test battery. To reduce practice effects, three different versions were provided for each test, and participants received different versions at consecutive visits. The three versions of all tests were constructed to be equally difficult and measure the same underlying cognitive function. Both test orders and test versions were divided equally between all participants. The battery included instruments that measured speed of processing, episodic memory, spatial ability, confidence judgments (Dahl, Allwood, Rennemark, & Hagberg, 2010), word comprehension (Dureman, 1960), and global cognitive function measured by MMSE (Folstein et al., 1975). Retest correlations for most of the tests were fairly high ($r = .51–.80$; median = .69), which indicates acceptable lower boundaries. However, for the test recognition ($r = .32$) and general knowledge ($r = .13$) the retest effect was low (Lövden et al., 2014). After testing, participants were interviewed about major life events, depression and coping strategies.

Depression was assessed by the Montgomery-Åsberg Depression Rating Scale. This scale was based on Comprehensive Psychiatric Rating Scale (CPRS; Montgomery & Asberg, 1979). The CPRS has repeatedly been used to assess depression in elderly populations (Bäckman, Hill, & Forsell, 1996; Pantzar et al., 2014). The scale consisted of ten questions regarding e.g. depression, anxiety, concentration and sleep deprivation. Each question had a scale from 0 to 6, with a total score of 60. According to Snaith, Harrop, Newby, and Teale (1986) scores of 0–6 indicated *free from depression*, 7–19 *mild depression*, 20–34 *moderate depression* and 35–60 *severe depression*.

2.3.1. Speed of processing

Two tests were included and designed to measure speed of processing, *the pattern comparison task* and the *digit cancellation task*. A structural equation model by Laukka et al. (2013) has found these two tests to be strongly correlated with each other. The pattern comparison task (Salthouse & Babcock, 1991) was a pencil and paper task with two separately administrated pages, consisting of 30 pairs of line-segment patterns, half of which were identical and half different, divided equally into two columns on each page. Participants had 30 s per page to identify as many identical pairs as possible. Three practice pattern-pairs were first completed to ensure comprehension of the task. Test score assessment was the total number of correctly classified patterns from both pages, maximum 60 correct patterns (Salthouse & Babcock, 1991). The digit cancellation task (Zazzo, 1974) was a pencil and paper task which originated from well-established cancellation tasks such as those used in Diller et al. (1974) and Lewis and Rennick (1979). The task consisted of one page with 11 rows of digits, ranging from 1 to 9, with one isolated practice row at the top of the page. There were 43 4's and the participant had 30 s to cross over as many of these as possible. Test score assessment was the total number of correctly crossed over digit "4".

2.3.2. Episodic memory

Two consecutive tests were designed to measure episodic memory, firstly *the free recall test* and secondly *the recognition test*. As for the speed of processing tests, a structural equation model showed that the free recall and recognition tests are strongly correlated. The test consisted of learning a word list, comprised of 16 unrelated concrete nouns (target words), with an immediate free recall trial and a forced yes/no recognition trial (Gardiner & Java, 1993). One at a time, words in a booklet were presented orally and visually for 5 s, in order to bypass visual and hearing problems, and to ensure registration of each word. Immediately after the presentation, participants were given two minutes for oral free recall. For the self-paced recognition task, the 16 target words intermixed with 16 interference words were presented in the same manner. Participants answered yes or no if they recognised a word or not. Of the interference words, four were perceptually similar, four were semantically similar and eight were unrelated to the target words. Following recognition of the word, participants were asked to choose between three alternatives. (1) If they remembered some specific detail ("remembered" item/source memory) from when the word was first presented to them. For example, whether they made an associative relation to the word, or inferred personal relevance to the word. (2) If they recognised the word as being familiar, but could not report a source memory for the word ("know" item based on familiarity). (3) If they were indecisive between alternative 1 and 2. Participants were requested to justify why they picked a specific alternative, confirming comprehension of the differences between remember/source memory and know/familiarity. Significant differences between the three versions have been found, participants performed somewhat better on version 1, compared to 2 and 3 in Laukka et al. (2013). Test score assessment for free recall was the total number of correctly recalled words, with a maximum of 16. For recognition scores, the recognised words were used, including both hits and false hits, combined into d-prime. d-prime is a concept originally stemming from signal detection theory, designed to distinguish between true hits and false hits (guessing), taking both hits and false hits into consideration (Snodgrass & Corwin, 1988).

2.3.3. Spatial ability

A *mental rotation test* (Rehnman & Herlitz, 2006) was designed to measure spatial ability. 10 series of 4 two-dimensional (quasi three-dimensional) figures of the Shepard and Metzler type (Shepard & Metzler, 1971) were presented in a booklet, one at a time. Each figure consisted of 10 cubes and there were three different versions of the 10 x 4 series. Each series contained one original figure placed to the left of three figures. One of the figures to the right was the same as the original figure but rotated to a specific angle. The two remaining figures were distractors, and could not be rotated to be congruent with the original figure. Participants were requested to choose between the three figures to the right and, as fast as possible, pick the correct rotated original figure. If the participant had not answered after 35 and 45 s, respectively, they were encouraged to do so. Three practice series were completed prior to the test in order to ensure comprehension. Test score assessment was divided into two elements, speed and accuracy. Speed was measured by seconds required for the participant to point at the correct figure.[2] Each task had a difficulty rating, which was the proportion of correct answers obtained when the test was assigned to a group of participants.[3] A difficult task endured a lower percentage of correct responses. Difficulty rating was considered when assessing speed response for each series (10 x participant). Two out of 30 tasks were missing difficulty ratings due to subsequent changes of the test after the difficulty ratings were produced, and therefore could not be used in the analysis. Accuracy was measured by number of correct responses, divided by the number of tasks in each version the participant had completed. The majority of participants completed all 10 tasks.

2.4. Statistical analyses

The data were analysed using a series of linear mixed models using IBM SPSS Statistics 22. The equation used for the mixed models was the following. The outcome (dependent variable of interest, i.e. the test score) y_{ijk} (corresponding to participant *i* meeting with test administrator *j* at visit *k*) is assumed to satisfy the relation:

$$y_{ijk} = \text{Prediction}_{ik} + \varepsilon_i + \varepsilon_j + \varepsilon_{ijk}.$$

The Prediction_{ik} represents nine predictors (independent variables as fixed effects) and the error terms ε_i, ε_j (random effects) and ε_{ijk} are assumed to follow independent zero mean Gaussian distributions with variances σ_1^2, σ_2^2 and σ_3^2, respectively. This assumption leads to the correlation $\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2+\sigma_3^2}$ between two measurements coming from one participant and $\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2+\sigma_3^2}$ between two measurements corresponding to one test administrator. There were in total five categorical predictors related to testing setting; time of day of the testing, the test order, versions of the test, testing setting. For speed of mental rotation, difficulty level of the task was included instead of test version. There were also three categorical predictors related to the participant; age group, sex, and previous experience of cognitive testing. Years of education of the participant was included as a continuous predictor. The testing specific factors were further assessed by estimated marginal means, adjusted for multiple testing using the Sidak-method. Cohen's *d* effect sizes were reported for each test (Cohen, 1992).

3. Results

3.1. Test administrator effects

There was a significant random effect of the test administrators for all the cognitive tests ($p < .01$), see Table 4 for detailed statistics. The variance components analysis revealed that the proportion of variance of the random intercept corresponding to test administrator was (.37/(.37 + 3.92 + 9.07)) = .028 for digit cancellation. This suggests that about 2.8%, of the total variation in the participant scores for digit cancellation, was ascribed to the test administrators, see Table 8 for further variance calculations. This entails that the total variance in the participants' test scores, determined by the factor corresponding to the test administrators, was between 1.4 and 3.5%.

3.2. Testing specific factor effects

Results showed that different testing specific factors were associated with particular cognitive tests (see Tables 5–8). When examining the estimated marginal means, significant differences between test versions were found for all the cognitive tests, except for accuracy for mental rotation. For digit cancellation, participants performed worse on version one in comparison to version three and for pattern comparison participants performed worse on version three in comparison to version two. Participants tested with version three of the episodic memory tests recalled significantly fewer words, and had poorer d'prime scores than those participants who received version one and two. Better performance scores were found when the speed of processing task was administrated as the second speed task in the test battery. A similar significant tendency was found for speed of mental rotation, i.e. having the mental rotation tasks further into the test battery was beneficial for a faster response, compared to having the mental rotation test at the very beginning of the test battery. Estimated marginal means showed that participants performed better on the recognition test in the morning testing sessions compared to before lunch, or in the afternoon and, in comparison to afternoon sessions, participants performed better before lunch. However, the reverse effect was seen for speed of mental rotation, where participants were slower when solving the task in morning compared to before lunch and in the afternoon. Participants that were tested at home performed worse on both speed of processing tasks and memory tasks compared to the participants that were tested at the testing centre. For accuracy of mental rotations scores, there were no significant effects of any of the testing specific factors contributing to participants' test scores.

Table 4. Variance estimates for different tests and variance components and percentage of variance corresponding to test administrator

Random effects	Variance estimate	Standard error	Wald *Z*	*p*-value	Percentage of variance
Digit cancellation					
Test administrator	.369	.129	2.86	.004**	2.8
Participant	9.07	.282	32.1	.001***	
Residual	3.92	.118	33.2	.001***	
Pattern comparison					
Test administrator	1.32	.437	2.82	.005**	3.5
Participant	22.3	.746	29.9	.001***	
Residual	11.7	.355	32.9	.001***	
Free recall					
Test administrator	.065	.029	2.26	.024*	1.4
Participant	1.92	.099	19.4	.001***	
Residual	2.59	.077	33.8	.001***	
Recognition					
Test administrator	.021	.008	2.52	.012*	2.6
Participant	.314	.017	18.4	.001***	
Residual	.473	.014	34.1	.001***	
Speed of mental rotation					
Test administrator	1.64	.573	2.68	.007**	1.5
Participant	27.8	.851	32.7	.001***	
Residual	71.6	.569	125.9	.001***	
Accuracy of mental rotation					
Test administrator	.0006	.0002	2.49	.013*	1.9
Participant	.008	.0007	11.6	.001***	
Residual	.024	.0007	34.5	.001***	

*Significant at the $p < .05$ level.

**Significant at the $p < .01$ level.

***Significant at the $p < .001$ level.

In addition, the analyses revealed poorer test performances with an increase in age (see Table 9). Length of education had a positive effect on all test scores. Moreover, women had better test scores than men on all cognitive tests, except for mental rotation, where men were both faster and had more correct responses. Previous experience of testing situation was favourable for tests scores of digit cancellation and speed of mental rotation. The correlation of test scores stemming from the same participants were large for all tests, varying between .248 and .680, suggesting that about 24.8–68% of the total variance in the participant's test scores was explained at the participant level.

Table 5. Statistics for time of day with estimated means, *F*-values and *p*-values for fixed effects and effect sizes of pairwise comparisons

Tests	Estimated marginal means† (*SE*)			*df*			*F*-value	*p*-value	Cohen's *d*		
	Morn.	P. lunch	Aftern.	Morn.	P. lunch	Aftern.			Morn. vs. P. lunch	Morn. vs. Aftern.	P. lunch vs Aftern.
Digit cancellation	15.3 (.172)	15.4 (.171)	15.5 (.173)	48.9	48.0	50.5	1.32	.268	.012	.014	.029
Pattern comparison	23.5 (.301)	23.6 (.3)	23.8 (.303)	40.8	49.1	42.2	1.54	.214	.081	.047	.029
Free recall	6.15 (.09)	6.11 (.89)	6.02 (.091)	68.1	66.2	73.4	1.81	.164	.021	.059	.038
Recognition	2.75 (.043)	2.67 (.43)	2.6 (.044)	46.9	45.8	49.4	14.4	.001***	.089	.168	.079
Speed of mental rotation	16.7 (.335)	16.2 (.333)	16.2 (.337)	36.4	35.8	37.6	6.28**	.002**	.046	.051	.006
Accuracy of mental rotation	.604 (.008)	.596 (.008)	.598 (.008)	62.3	60.4	67.3	1.28	.278	.045	.033	.011

Notes: Morn. = Morning, P. lunch = Prior lunch, Aftern. = Afternoon.

†Estimated means are evaluated at 10.9 years of education for all tests except speed of mental rotation which is evaluated at 11.14 years.

**Significant at the *p* < .01 level.

***Significant at the *p* < .001 level.

Table 6. Statistics for test order with estimated means, *F*-values and *p*-values for fixed effects and effect sizes of pairwise comparisons

Tests	Estimated marginal means† (*SE*)		*df*		*F*-value	*p*-value	Cohen's *d*
	Order 1	Order 2	Order 1	Order 2			Order 1 vs. 2
Digit cancellation	15.3 (.169)	15.6 (.17)	46.1	46.8	12.4	.001***	.082
Pattern comparison	24.3 (.297)	23.0 (.298)	38.6	39.0	87.8	.001***	.226
Free recall	6.07 (.086)	6.11 (.087)	59.1	60.4	.425	.515	.017
Recognition	2.66 (.042)	2.69 (.42)	41.9	42.6	1.99	.158	.036
Speed of mental rotation	15.8 (.331)	16.9 (.331)	34.9	35.0	42.8	.001***	.010
Accuracy of mental rotation	.598 (.008)	.6 (.008)	54.15	54.2	.228	.633	.011

†Estimated means are evaluated at 10.9 years of education for all tests except speed of mental rotation which is evaluated at 11.14 years.

***Significant at the *p* < .001 level.

Table 7. Statistics for test versions with estimated means, *F*-values and *p*-values for fixed effects and effect sizes of pairwise comparisons

Tests	Estimated marginal means[†] (*SE*)			*df*			*F*-value	*p*-value	Cohen's *d*		
	Version 1	Version 2	Version 3	Version 1	Version 2	Version 3			Version 1 vs. 2	Version 1 vs. 3	Version 2 vs. 3
Digit cancellation	15.5 (.169)	15.4 (.17)	15.3 (.17)	46.4	46.8	47.0	6.02	.001***	.039	.071	.034
Pattern comparison	23.6 (.297)	23.9 (.298)	23.4 (.298)	39.0	39.4	39.4	8.52	.001***	.052	.038	.09
Free recall	6.3 (.088)	5.8 (.088)	6.17 (.88)	64.8	64.8	64.8	42.2	.001***	.234	.06	.173
Recognition	2.65 (.043)	2.72 (.043)	2.66 (.043)	45.0	45.1	45.2	5.24**	.005**	.078	.007	.071
Speed of mental rotation	N/A	N/A	N/A	N/A	N/A	N/A	2,236	.001***	N/A	N/A	N/A
Accuracy of mental rotation	.593 (.008)	.605 (.008)	.599 (.008)	60.0	60.6	60.6	2.38	.93	.061	.033	.033

Note: The covariate difficulty level is only relevant for mental rotation speed, all remaining tests have test version as covariate instead.

[†] Estimated means are evaluated at 10.9 years of education for all tests except speed of mental rotation which is evaluated at 11.14 years.

**Significant at the *p* < .01 level.

***Significant at the *p* < .001 level.

Table 8. Statistics for testing setting with estimated means, *F*-values and *p*-values for fixed effects and effect sizes of pairwise comparisons

Tests	Estimated marginal mean[†] (*SE*)		*df*		*F*-value	*p*-value	Cohen's *d*
	Test centre	Home visit	Test centre	Home visit			Test centre vs. home visit
Digit cancellation	16.1 (.159)	14.8 (.204)	36.1	96.6	64.9	.001***	.356
Pattern comparison	24.9 (.28)	22.4 (.354)	30.8	77.51	87.2	.001***	.43
Free recall	6.34 (.077)	5.85 (.118)	38.1	199.1	19.2	.001***	.173
Recognition	2.74 (.039)	2.61 (.054)	30.0	114.9	6.87	.001***	.139
Speed of mental rotation	16.2 (.313)	16.5 (.394)	27.9	69.9	1.23	.268	.034
Accuracy of mental rotation	.598 (.007)	.6 (.011)	35.5	174.5	.862	.011*	.862

[†] Estimated means are evaluated at 10.9 years of education for all tests except speed of mental rotation which is evaluated at 11.14 years.

*Significant at the *p* < .05 level.

***Significant at the *p* < .001 level.

Table 9. *F*-values and *p*-values for fixed effects for individual differences

Tests	Individual differences							
	Age		Sex		Education		Previous experience	
	F-value	*p*-value	*F*-value	*p*-value	*F*-value	*p*-value	*F*-value	*p*-value
Digit cancellation	107.1	.001***	44.2	.001***	97.66	.001***	9.8	.002**
Pattern comparison	192.2	.001***	19.7	.001***	255.1	.001***	3.45	.064
Free recall	46.5	.001***	225	.001***	169.5	.001***	.53	.467
Recognition	3.53	.001***	78.4	.001***	66.5	.001***	.177	.674
Speed of mental rotation	31.9	.001***	5.69	.001***	31.36	.001***	22.6	.001***
Accuracy of mental rotation	16.7	.001***	320	.001***	104	.001***	1.37	.242

**Significant at the $p < .01$ level.

***Significant at the $p < .001$ level.

4. Discussion

The aims of this study were to investigate the influence of the test administrator and testing specific factors on test scores in a longitudinal study. Results showed a significant amount of test administrator influence on participant's test scores, where 1.4–3.5% of the variations in test scores were explained on a test administrator level. Influences of other testing specific factors, such as test version, test order, time of day and testing setting were found to significantly influence the majority of test scores measuring different cognitive functions.

4.1. Test administrator influence

In regard to previous research on underlying causes of test administrator influence, it is likely that the test administrator influence found here is partially related to experimenter variables (Sattler & Theye, 1967), e.g. experience, age and sex of the test administrator. This study included female and male test administrators, with large variation in testing experience (42–1,503 participants) and age range (25–50 years). Inexperience of test administrators has previously explained discrepancies in test scores (Hoyt & Kerns, 1999). However, separate analyses involving only experienced test administrators also showed a test administrator effect (data not shown). It thus seems reasonable to conjecture that experience of the test administrator does not account for all test administrator influence on test scores in our sample of test administrators. Although there was a large age discrepancy between test administrators, 71% of the administrators were between 25 and 33 (median 30), which is a rather small variation in age, making it difficult to draw conclusions of whether the age of the test administrator was an influential factor or not. Similarly, with two male and 18 female test administrators, we cannot assess the influence of the sex of the test administrator.

Additionally, we cannot rule out that the experimenter variables interact with other factors, e.g. concerning the characteristics of the participant. For instance, Ortner and Vormittag (2011) found that female participants performed better on verbal general knowledge test when tested by a female test administrator.

Sattler and Theye (1967) discuss that departures from standardised procedures, are more prominent in ageing studies than in studies using younger participants. Level of overall cognition declines with advancing age (Salthouse, 2009), and therefore samples of elder participants may require supplementary test instructions to comprehend the task at hand. This can lead to departures from standardised testing instructions when testing older participants. One test administrator might handle the elder's incomprehension differently from its colleague, and therefore systematic test administrator influences occur more often in studies examining elderly samples, compared to those focusing on younger samples. Future research could focus on test administrator influence in relation to the age of the sample, bringing insight to whether ageing studies should routinely inspect for test administrator influence.

Furthermore, differences in the magnitude of test administrator influence were observed between different types of tests. The largest proportion of the total variance attributed to test administrator was found for the speed of processing instruments and recognition, the smallest for mental rotation and free recall. The fact that the variance attributed to test administrator is positive for all tests may indicate that regardless of the type of test and what underlying cognitive function it is intended to capture, test administrator influence may exist in test scores.

The extent of test administrator influences on cognitive test scores in longitudinal studies has not been studied. Reasons for this include that rigid routines of test administration are usually applied, to purposely eliminate test administrator influence. Likewise, because of the complicated nature of test administrators influences and their underlying causes, investigators may struggle to comprehend to what extent test administrator influences exist in their data (Hoyt, 2000), leading to a hesitation in exploring such influences. Although this paper can only speculate why there is administrator influence on test scores, it demonstrates that despite rigid routines, variability in participants' performance is attributable to the test administrator, accounting for up to 4% of the variance.

4.2. Testing specific influences

There were four major findings of influences attributable to the testing specific factors. Firstly, no negative effect of the second test order was found, implying that serial testing was not mentally tiring for the participants. One could also then presume that our test battery was not strenuous for our participants. On the contrary, participants performed better on the speed of processing and the speed of mental rotation task further into the testing battery, signifying *task familiarity* or *immediate practice effects*. It has previously been suggested that participants perform better on tasks that they have already completed, because they fully comprehend what the test requires the second time they receive a similar test (Goldberg, Harvey, Wesnes, Snyder, & Schneider, 2015).

Secondly, there were effects of time of day for the recognition test, and best performances were found for morning sessions. Additionally, participants tested in the afternoon and prior lunch performed better than participants tested in the morning on the speed of mental rotation test. Differences in test scores ascribed to time of day have previously been explained by circadian rhythm influences on cognition, and participants tested at matched peak circadian periods show better cognitive performances than those tested at off-peak time of day (Blatter & Cajochen, 2007). These results have been found for a wide range of cognitive tasks, measuring different types of memory, executive functioning, and attention (Schmidt et al., 2007). However, not all types of performances on cognitive tasks are influenced equally (Blatter & Cajochen, 2007). Differences in task duration, cognitive load (high vs. low), and difficulty level could explain why we found specific time of day effects exclusively on recognition and speed of mental rotation, and not for the remaining tests. For example, less cognitively demanding tasks (e.g. the speed of processing tasks used here) have been found to be less affected by influences of circadian rhythms (Yoon, May, & Hasher, 1999). In regard to our elderly sample, May et al. (2005) also found that morning testing sessions, compared to testing sessions later in the day, were beneficial for performances on instruments measuring memory for individuals between 60 and 75 years old.

Thirdly, there were significant differences between test versions of all the cognitive tests except for accuracy of mental rotation. Differences between test versions of the episodic memory test have previously been presented in SNAC-Kungsholmen study (Laukka et al., 2013), and explained by variations in difficulty levels. Researchers that use different versions of the same instrument should routinely inspect and report methodological diversities, in order to reduce reactive effects.

Fourthly, participants tested in their home environment performed worse than those tested in the test centre. Many distractions may exist in a home environment influencing the scores negatively, such as uncontrollable background noise, distractions from home care assistance and so forth. Yet, these results are inconsistent with previous results from Shievitz et al. (1998), who found that geriatric participants tested at home scored significantly better on the MMSE than when tested in a clinical setting. Selection bias can provide insight to these contradictory results. Participants that receive home visits are those individuals that are not well enough to come to the testing centre. Although we applied a rather strict exclusion criterion, a selection of the most unfit may still remain in the home visit group, leading to poorer scores in this group in comparison to the testing centre group.

In sum, testing specific factors can be relevant for most cognitive performances, however future research is needed to confirm these results.

4.3. Limitations and strengths

One limitation in our study was not to include any potential influencing factors related to experimenter variables (characteristics of the test administrator) in the statistical analyses, making it impossible to draw conclusions regarding the impact of these factors. However, the main aim of this study was to solely examine the extent to which test administrator influence could exist in a longitudinal ageing study. Moreover, further potentially relevant information concerning the characteristics of test administrator, such as ethnicity or personality traits, was not accessible in our study. Neither was information regarding situational variables, e.g. praise or discouragement of the participant's performance during testing, and if this was used systematically by particular test administrators. Another weakness was not inspecting interactions of testing specific factors. For example, the effect on test score of a difficult test version may be different in the morning compared to the afternoon.

Strengths of the study include using a large random sample from the general population, covering both urban and rural areas, with multiple follow-ups and including diverse cognitive testing instruments. This increases generalisability of the results. Furthermore, by examining test administrator influence through a random effect, the sample of test administrators are statistically assumed to come from a hypothetical population of test administrators, which could also increase generalisability (Searle, 1971). Another strength, are our procedures of data collection, enabling an ability to measure and evaluate specific confounding factors. For example, two different orders of the test battery allow us to investigate fatigue effects due to serial testing on particular tests. Also, by documenting when the participant was tested, investigation of fatigue due to the time of day could be carried out.

Reducing test administrator bias on test scores is crucial in order to acquire internal validity, and ensure reliable cognitive instruments. High quality training of administrators, and good standardisation, of the instruments reduce these types of weaknesses (Rousson et al., 2002). Our study applied these types of precautions, such as training, extensive test manuals and regular meetings between test administrators. Furthermore, this study used acceptably reliable instruments. Still, there were differences in test scores attributable to the test administrators. It is however plausible that the effects are relatively small due to the very fact that rigid routines were applied. Therefore, it seems worthwhile to continue with these types of practices. Moreover, investigators should examine their data for systematic test administrator effects, alternatively use computer-based instruments, when appropriate. On a further note, systematic test administrator influences in cross-sectional data could also be of a concern, and our results could be applicable to cross-sectional studies using multiple test administrators to collect data.

When considering what type of procedural choices to implement, and practicalities to take into account, it is important to consider their impact on the outcome. For example, the results indicate that there were no effects of fatigue due to serial testing, therefore two orders of the testing battery could be excessive. The time of day was seen to influence test scores and therefore future studies should consider whether it is possible to test all participants at the same time of day. Comparing results from participants tested at home to those tested at the testing centre is also a challenge. Weighing the benefits of reducing selection bias against the potential influence of differences in testing environment warrants further investigation.

4.4. Concluding remarks

This study has provided evidence that the practice of using multiple test administrators to collect cognitive data is a tangible concern in longitudinal studies of ageing. The effects of test administrator influence on test scores are noteworthy, even though the magnitude of these effects is relatively small. We found that time of day, serial testing, different versions of the same test, and testing setting can also cause implications in interpretation of test scores. In summary, these results endorse Guo's (2013) assertion that researchers must consider both individual change and variation in testing settings when analysing longitudinal data.

Funding
The authors received no direct funding for this research.

Competing Interests
The authors declare no competing interest.

Author details
Marieclaire Overton[1]
E-mail: marie_claire.overton@med.lu.se
ORCID ID: http://orcid.org/00-0001-6708-6273
Mats Pihlsgård[1]
E-mail: mats.pihlsgard@med.lu.se
ORCID ID: http://orcid.org/0000-0002-3934-9387
Sölve Elmståhl[1]
E-mail: solve.elmstahl@med.lu.se
ORCID ID: http://orcid.org/0000-0001-7153-5414
[1] Division of Geriatric Medicine, Lunds University, Skånes University Hospital, Jan Waldenströms gata 35, CRC Building 28, fl.13, SE-205 02 Malmö, Sweden.

Citation information
Cite this article as: Test administrator effects on cognitive performance in a longitudinal study of ageing, Marieclaire Overton, Mats Pihlsgård & Sölve Elmståhl, *Cogent Psychology* (2016), 3: 1260237.

Notes

1. Other testing specific factor may exist than the ones mentioned here, these are relevant for our study.
2. That the time it takes for the participant to point at the incorrect figure is irrelevant here.
3. Over 200 students in a pilot study, who responded to 20 tasks each, of all three versions.

References

Bäckman, L., Hill, R. D., & Forsell, Y. (1996). The influence of depressive symptomatology on episodic memory functioning among clinically nondepressed older adults. *Journal of Abnormal Psychology, 105*, 97–105. doi:10.1037/0021-843X.105.1.97

Blatter, K., & Cajochen, C. (2007). Circadian rhythms in cognitive performance: Methodological constraints, protocols, theoretical underpinnings. *Physiology & Behavior, 90*, 196–208. doi:10.1016/j.physbeh.2006.09.009

Bookout, D. V., & Hosford, R. E. (1969). Administration effects on the S-329 of the GATB using three experimental treatments. *Journal of Employment Counseling, 6*, 124–133. doi:10.1002/j.2161-1920.1969.tb00523.x

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. doi:10.1037/0033-2909.112.1.155

Dahl, M., Allwood, C., Rennemark, M., & Hagberg, B. (2010). The relation between personality and the realism in confidence judgements in older adults. *European Journal of Ageing, 7*, 283–291. doi:10.1007/s10433-010-0164-2

Diller, L., Ben-Yishay, Y., Gerstman, L. J., Goodkin, R., Gordon, W., & Weinberg, J. (1974). *Studies in cognition and rehabilitation in hemiplegia* (Rehabilitation Monograph No.50). New York, NY: New York University Medical Center.

Dureman, I. (1960). *SRB: 1*. Stockholm: Psykologiförlaget.

Ekström, H., & Elmståhl, S. (2006). Pain and fractures are independently related to lower walking speed and grip strength: Results from the population study "Good Ageing in Skåne". *Acta Orthopaedica, 77*, 902–911. doi:10.1080/17453670610013204

Ferrer, E., & Ghisletta, P. (2011). Methodological and analytical issues in the psychology of ageing. In W. Schaie & S. L. Willis (Eds.), *Handbook of the psychology of ageing* (7th ed., pp. 25–39). London: Academic Press. http://dx.doi.org/10.1016/B978-0-12-380882-0.00002-4

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state". *Journal of Psychiatric Research, 12*, 189–198. doi:10.1016/0022-3956(75)90026-6

Gardiner, J. M., & Java, R. I. (1993). Recognizing and remembering. In A. F. Collins, S. E. Gathercole, M. A. Conway, & P. E. Morris (Eds.), *Theories of memory* (pp. 163–188). Hove: Erlbaum.

Goldberg, T. E., Harvey, P., Wesnes, K. A., Snyder, P. J., & Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: Impliactions for preclinical Alzheimer's disease randomized controlled trials. *Cognitive & Behavioral Assessment, 1*, 103–111. doi:10.1016/j.dadm.2014.11.003

Guo, S. (2013). Correction of rater effects in longitudinal research with a cross-classified random effects model. *Applied Psychological Measurement, 38*, 37–60. doi:10.1177/0146621613488821

Guo, S., & Bollen, K. A. (2013). Research using longitudinal ratings collected by multiple raters: One methodological problem and approaches to its solution. *Social Work Research, 37*, 85–98. doi:10.1093/swr/svs029

Hofer, M. S., & Schaie, K. W. (2001). Longitudinals studies in ageing research. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of ageing* (5th ed., pp. 53–77). San Diego, CA: Academic Press.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*, 64–86. doi:10.1037/1082-989X.5.1.64

Hoyt, W. T., & Kerns, M. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*, 403–424. doi:10.1037/1082-989X.4.4.403

Lagergren, M., Fratiglioni, L., Hallberg, I. R., Berglund, J., Elmståhl, S., Hagberg, B., ... Wimo, A. (2004). A longitudinal study integrating population, care and social services data. The Swedish national study on ageing and care (SNAC). *Ageing Clinical Experimental Research, 16*, 158–168. http://dx.doi.org/10.1007/BF03324546

Laukka, E. J., Lövdén, M., Herlitz, A., Karlsson, S., Ferencz, B., Pantzar, A., ... Bäckman, L. (2013). Genetic effects on old-age cognitive functioning: A population-based study. *Psychology and Ageing, 28*, 262–274. doi:10.1037/a0030829

Lewis, R. F., & Rennick, P. M. (1979). *Manual for the repeatable cognitive-perceptual-motor battery*. Grosse Pointe Park, MI: Axon.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*, 543–560. doi:10.1177/0265532211406422.

Lissner, L. M., Skoog, I., Andersson, K., Beckman, N., Sundh, V., Waern, M., ... Björkelund, C. (2003). Participation bias in longitudinal studies: Experience from the population study of women in Gothenburg, Sweden. *Scandinavian Journal of Primary Health Care, 21*, 242–247. doi:10.1080/02813430310003309-1693

Lövden, M., Köhnke, Y., Laukka, E. J., Kalpouzos, G., Salami, A., Li, T. Q., ... Bäckman, L. (2014). Changes in perceptual speed and white matter microstructure in the corticospinal tract are associated in very old age. *Neuroimage,* 15, 520–530. doi:10.1016/j.neuroimage.2014.08.020

Marx, D. M., & Goff, P. A. (2005). Clearing the air: The effect of experimenter race on target's test performance and subjective experience. *British Journal of Social Psychology, 44*, 645–657. doi:10.1348/014466604X17948

May, C. P., Hasher, L., & Foong, N. (2005). Implicit memory, age, and time of day. *Psychological Science, 16*, 96–100. doi:10.1111/j.0956-7976.2005.00788.x

Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry, 134*, 382–389. doi:10.1192/bjp.134.4.382

Ortner, T. M., & Vormittag, I. (2011). Test administrator's gender affects female and male students' self-estimated verbal general knowledge. *Learning and Instruction, 21*, 14–21. doi:10.1016/j.learninstruc.2009.09.003

Pantzar, A., Laukka, E. J., Atti, A. R., Fastbom, J., Fratiglioni, L., & Bäckman, L. (2014). Cognitive deficits in unipolar old-age depression: A population-based study. *Psychological Medicine,* 44, 937–947. doi:10.1017/S0033291713001736

Rehnman, J., & Herlitz, A. (2006). Higher face recognition ability in girls: Magnified by own-sex and own-ethnicity bias. *Memory, 14*, 289–296. doi:10.1080/09658210500233581

Rosenthal, R., & Fode, K. L. (1963). Psychology of the scientist: V. Three experiments in experimenter bias monograph supplement 3-V12. *Psychological Reports, 12*, 491–511. doi:10.2466/pr0.1963.12.2.491

Rousson, V., Gasser, T., & Seifert, B. (2002). Assessing intrarater, interrater and test-retest reliability of continuous measurements. *Statistics in Medicine, 21*, 3431–3446. doi:10.1002/sim.1253

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of Ageing, 30*, 507–514. doi:10.1016/j.neurobiolageing.2008.09.023

Salthouse, T. A. (2014). Ageing cognition unconfounded by prior test experience. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 65*, 698–705. doi:10.1093/geronb/gbu063

Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology, 27*, 763–776. doi:10.1037/0012-1649.27.5.763

Samuel, W. (1977). Observed IQ as a function of test atmosphere, tester expectation, and race of tester: A replication for female subjects. *Journal of Educational Psychology, 69*, 593–604. doi:10.1037/0022-0663.69.5.593

Sattler, J. M., & Theye, F. (1967). Procedural, situational, and interpersonal variables in individual intelligence testing. *Psychological Bulletin, 68*, 347–360. doi:10.1037/h0020194

Schmidt, C., Collette, F., Cajochen, C., & Peigneux, P. (2007). A time to think: Circadian rhythms in human cognition. *Cognitive Neuropsychology, 24*, 755–789. doi:10.1080/02643290701754158

Schmidt, K. R. T., & Teti, D. M. (2005). Issues in the use of longitudinal and cross-sectional designs. In D. M. Teti (Ed.), *Handbook of research methods in developmental science* (pp. 3–20). Malden, MA: Blackwell Publishing.

Searle, S. R. (1971). *Linear models*. New York, NY: Wiley.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Harcourt.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*, 701–703. doi:10.1126/science.171.3972.701

Shievitz, A. L., Tudiver, F., Araujo, A., Sanghe, P., & Boyle, E. (1998). Do elderly people score better on cognitive tests at home? *Canadian Familiy Physician, 44*, 1652–1656.

Snaith, R., Harrop, F., Newby, D., & Teale, C. (1986). Grade scores of the Montgomery-Asberg depression and the clinical anxiety scales. *The British Journal of Psychiatry, 148*, 599–601. doi:10.1192/bjp.148.5.599

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology, 1*, 35–50. doi:10.1037/0096-3445.117.1.34

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25–29. doi:10.1037/h0071663

Yoon, C., May, C. P., & Hasher, L. (1999). Ageing, circadian arousal patterns, and cognition. In N. Schwarz, D. Park, B. Knauper, & S. Sudman (Eds.), *Ageing, cognition and self reports* (pp. 117–143). Washington, DC: Psychological Press.

Zazzo, R. (1974). *Test des deux barrages. Actualités pédagogiques et psychologiques [Test of the two dams. News pedagogical and psychological]* (Vol. 7). Neuchâtel: Delachaux et Nestlé.

***Cogent Psychology* (ISSN: 2331-1908) is published by Cogent OA, part of Taylor & Francis Group.**