Vision Research 215 (2024) 108355

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Weak correlations between visual abilities in healthy older adults, despite long-term performance stability

Simona Garobbio [a,*], Marina Kunchulia [b], Michael H. Herzog [a]

[a] Laboratory of Psychophysics, Brain Mind Institute, EEcole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[b] Free University of Tbilisi and Ivane Beritashvili Center of Experimental Biomedicine, Tbilisi, Georgia

ARTICLE INFO

Keywords:
Visual common factor
Test-retest reliability
Longitudinal study
Battery of visual tests
Healthy aging
Individual differences

ABSTRACT

Using batteries of visual tests, most studies have found that there are only weak correlations between the performance levels of the tests. Factor analysis has confirmed these results. This means that a participant excelling in one test may rank low in another test. Hence, there is very little evidence for a common factor in vision. In visual aging research, cross-sectional studies have repeatedly found that healthy older adults' performance is strongly deteriorated in most visual tests compared to young adults. However, also within the healthy older population, there is no evidence for a visual common factor. To investigate whether the weak between-tests correlations are due to fluctuations in individual performance throughout time, we conducted a longitudinal study. Healthy older adults performed a battery of eight visual tests, with two re-tests after approximately four and seven years. Pearson's, Spearman's and intraclass correlations of most visual tests were significant across the three testing, indicating that the tests are reliable and individual differences are stable across years. Yet, we found low between-tests correlations at each visit, which is consistent with previous studies finding no evidence for a visual common factor. Our results exclude the possibility that the weak correlations between tests are due to high within-individual variance across time.

1. Introduction

The idea of common mechanisms is encountered in everyday life. For example, the factor '*g*' is claimed to represent a common factor for intelligence, which can be inferred from several specific factors, such as the Wechsler scale (Wechsler, 2003). Similarly, there is evidence of a high correlation between touch and audition, suggesting a common factor for somatosensation (Frenzel et al., 2012). Likewise, it is reasonable to expect a common factor for vision. For instance, in order to obtain a driving license, we need to pass a visual acuity test (Owsley & McGwin, 2010). The result of this test is often considered a gold standard that represents general visual abilities. If indeed there is a common factor underlying visual abilities, performance in various visual tests should strongly correlate with each other.

Cappe and colleagues (2014) used a visual test battery to investigate the implicit assumption of such a visual common factor. Forty young participants performed six basic visual tests, of which five were spatial vision tests, such as the Freiburg visual acuity test and a vernier offset discrimination task. Contrary to expectation, only weak correlations were found between the performances, suggesting no evidence for a visual common factor. Most other studies have reported very similar results (e.g., Bargary et al., 2017; Bosten & Mollon, 2010; Cretenoud et al., 2019; Grzeczkowski, Clarke, Francis, Mast, & Herzog, 2017; Verhallen et al., 2017; for reviews see Bosten, Mollon, Peterzell, & Webster, 2017; Peterzell, 2016; Tulver, 2019).

The aforementioned studies mainly involved young participants. Since the age-related decline of ophthalmological and cortical factors, such as lens clouding and reduction of neurons, affect some people more than others, one might expect a stronger visual common factor in healthy older adults. Cross-sectional studies have repeatedly shown that healthy older adults have deteriorated performance in most visual tests compared to young adults (e.g., visual search Scialfa, Esau, & Joffe, 1998; contrast sensitivity Delahunt, Hardy, & Werner, 2008; visual backward masking Plomp, Kunchulia, & Herzog, 2012; motion detection Bocheva, Angelova, & Stefanova, 2013). These visual tests are (implicitly) believed to target the relevant mechanisms of visual decline associated with aging. Consequently, an older adult who is more affected by age-related decline than other adults, is expected to perform worse in all visual tests. To test it, Shaqiri et al., (2019) assessed performance of 104 young and 92 healthy older participants in 16 visual tests spanning

* Corresponding author at: Laboratory of Psychophysics, Brain Mind Institute, School of Life Sciences, EFPL, CH-1015 Lausanne, Switzerland.
E-mail address: simona.garobbio@epfl.ch (S. Garobbio).

https://doi.org/10.1016/j.visres.2023.108355
Received 30 June 2023; Received in revised form 11 December 2023; Accepted 11 December 2023
Available online 23 December 2023

many visual abilities. In each visual test, younger adults performed significantly better, on average, than healthy older adults. However, between-tests correlations were very weak in both groups, suggesting that poor performance in one visual test by an older adult does not necessarily indicate poor performance in another test. Therefore, even in healthy aging, there appears to be no evidence for a common factor underlying visual abilities.

There are several potential explanations for the weak between-tests correlations observed. One possibility is that the tests themselves may have low test–retest reliability. However, previous studies have shown rather high short-term test–retest reliability, with Pearson's or intraclass correlations from approximately 0.50 to 0.90 (Cappe et al., 2014; Cretenoud et al., 2019; Grzeczkowski et al., 2017; Shaqiri et al., 2019). Another possibility, which we investigate in this study, is that long-term performance stability is low due to within-individual fluctuations throughout time. In this study, we refer to short-term test–retest reliability when the test-retests were carried out closely in time, capturing measurement errors, and to long-term performance stability when retests were performed years after the initial testing, capturing changes in participants' performance. These changes in participants' performance may occur due to the impact of healthy aging and/or fluctuations in our visual abilities caused by the dynamic nature of our brains. To explore this further, we conducted a longitudinal study to examine the long-term performance stability of various visual tests. We invited the older adults in the study by Shaqiri et al., (2019) to perform eight visual tests again after approximately four and seven years from the first testing. These visual tests covered various visual functions, including contrast sensitivity, spatial vision (vernier offset discrimination, visual acuity, and orientation discrimination), motion perception, visual search, and speed (reaction time and Simon tests). The tests were selected because strong correlations between tests within the same visual function were expected, and it was hypothesized that both contrast sensitivity and spatial vision play a role (in a hierarchical manner) to all of these visual functions.

2. Methods and Materials

2.1. Participants

Older adults were invited for three visits at zero, four and seven years apart to the Beritashvili Center of Experimental Biomedicine in Tbilisi, Georgia. Out of the 92 older adults included in the first visit (published in Shaqiri et al., 2019), 61 returned for the second visit, and 39 for the third visit. We excluded participants who developed eye disease after the first visit, such as maculopathies (n = 3 in visit 2 and n = 1 in visit 3) and glaucoma (n = 1 in visit 3). Thus, included participants had no known history of eye disease (e.g., traumatic injury, thrombosis, glaucoma or maculopathies) nor, based on participants' report, any known diagnosis of dementia, Parkinson's disease, sequels to brain injury (e.g., trauma or stroke), or any other disorder known to affect visual or cognitive abilities. Please note that cataract (n = 8 in visit 2, n = 7 in visit 3) was not an exclusion criterion because the participants' visual acuity fell within the range of the one of the other participants, and excluding them did not alter the interpretation of the results (data not shown).

All results are reported for two groups: the 58 healthy older adults who participated in the firsts two visits (age visit 1 = 64.2 ± 3.52; age visit 2 = 68.1 ± 3.62; 39 female; education degree: compulsory school n = 2, high school n = 9, university n = 46, all right-handed), and the 37 healthy older adults who participanted in all three visits (age visit 1 = 64.1 ± 3.61; age visit 2 = 68.1 ± 3.71; age visit 3 = 70.8 ± 3.73; 24 female; education degree: compulsory school n = 1, high school n = 3, university n = 32, all right-handed). A description of participants' medical history including refractive errors is provided in Table 1 for the group of older adults participating in visits 1 and 2, and in Table S1 in the Supplementary Material for the group of older adults participating in visits 1, 2 and 3.

The study complied with the Declaration of Helsinki and was approved by the ethics committee of the Beritashvili Center of Experimental Biomedicine in Tbilisi, Georgia. All participants provided written informed consent, were reimbursed for their participation, and were informed that they could withdraw from the experiment at any time. The experimental sessions of each visit occurred on two consecutive days and lasted approximately 60 min each. The order of the tests was randomized across participants and visits. Data from the complete set of older adults tested at visit 1 were published in two previous studies (Garobbio, Pilz, Kunchulia, & Herzog, 2022; Shaqiri et al., 2019). Please note that only older adults tested in Georgia were re-invited, and a subset of tests was used for the longitudinal study.

2.2. General methods and apparatus

During each visit, all participants performed a battery of eight visual tests. The battery covered different visual functions: vernier offset discrimination, visual acuity, orientation discrimination, contrast sensitivity, motion direction sensitivity, simple reaction time, visual search, and the Simon test. Some participants did not complete all tests for various reasons (e.g., prematurely quitting the experiment), but they were not excluded from the dataset (see Section 2.4.1).

For visit 1, stimuli were displayed on a Samsung SyncMaster 957DF CRT monitor (31 cm × 23 cm, 1024 × 768 pixels, 100 Hz). Due to technical issues, an ASUS VG248QE LCD monitor (53 cm × 30 cm, 1920 × 1080 pixels, 120 Hz) was used for visits 2 and 3, while efforts were made to maintain spatial and temporal stimulus properties as comparable as possible. Specifically, to account for the faster pixel onset of the CRT in comparison to the LCD, 2 ms were added to the reaction times in the simple reaction time test and visual search test for visit 1. The monitors were calibrated to output a maximum white luminance of 80 cd/m^2. Participants sat in a dimly illuminated room and, when applicable, were instructed to wear their glasses. Participants sat 5 m away from the screen for the visual acuity and the vernier offset discrimination tests, while for all other tests, the distance to the screen was 2 m.

An auditory feedback tone was provided after incorrect responses in all visual tests except for the visual acuity and the simple reaction time tests. Unless otherwise stated, participants used hand-held push buttons

Table 1
Description of refractive errors, medication and medical follow-up of the 58 participants participating in visits 1 and 2.

Refractive errors	visit1% (n)	visit2% (n)	Medication	visit1% (n)	visit2%, (n)	Medical follow-up	visit1% (n)	visit2% (n)
Myopia	5.17 (3)	3.45 (2)	None	50.0 (29)	37.9 (22)	None currently	86.2 (50)	82.8 (48)
Presbyopia	60.3 (35)	56.9 (33)	Hypotensors & statins	25.9 (15)	44.8 (26)	Check-up only	None	None
Both myopia & presbyopia	32.8 (19)	37.9 (22)	GERD & heartburn drugs	None	5.17 (3)	For rheumatic disorders	3.45 (2)	None
Emmetropia	1.7 (1)	1.7 (1)	Chondrosulfites	None	None	For cardio-vascular disorders	5.17 (3)	6.90 (4)
			Lorazepam (BZD)	None	None	For other disorders[a]	5.17 (3)	12.1 (7)
			Anti-histaminic	1.74 (1)	None			
			Anti-epileptic	None	None			
			Others[a]	29.3 (17)	31.0 (18)			

[a] None of which are known to affect vision nor cognition.

to provide their responses. Thresholds were measured using an adaptive PEST procedure (Taylor & Creelman, 1967), aiming for a 75 % correct response rate in 2-alternative forced choice test (2AFC). However, the Freiburg visual acuity adopted the Best-PEST adaptive procedure, aiming for a 62.5 % correct response rate in a 4-alternative forced choice test (4AFC). The stimulus programs were implemented in C/C++ using a stimulus presentation library developed in-house.

2.3. Visual tests

Vernier offset discrimination (VO): A vernier stimulus consisting of two vertical bars slightly offset in the horizontal direction (Fig. 1a) was presented for 150 ms with a random offset direction. Participants were required to indicate the offset direction of the lower bar in relation to the upper bar (left vs. right). The vernier offset was adaptively varied, and the offset threshold (VO) was determined. Please note that despite the big observation distance, vernier offsets were limited by the monitors pixel resolution to an integer multiple of about 12 arcsec. This was still sufficiently good to collect enough information for getting a stable fit of the psychometric function from which the threshold could be extracted.

Freiburg visual acuity (VA): The Freiburg visual acuity test (Bach, 1996) was utilized to measure binocular visual acuity (VA). Landolt-C optotypes with a gap in one of four possible orientations ("up", "down", "left" or "right") were presented (Fig. 1b). The size of the optotype changed adaptively, and the orientation of the gap was randomly selected. Participants were instructed to verbally indicate the direction of the gap, and the experimenter operated the input device accordingly. The test provided the decimal visual acuity value.

Orientation discrimination (Ori): The test was based on the one used by Tibber et al. (2006). Participants were instructed to determine whether a Gabor patch, whose orientation changed adaptively, was oriented clockwise or counterclockwise relative to the vertical axis. The Gabor patch had a Michelson contrast of 80 %, mean luminance of 40 cd/m^2, spatial frequency of 3.3 cycles/arcdeg, an envelope sigma along the orientation of 0.57 arcdeg, and an envelope sigma perpendicular to the orientation of 0.19 arcdeg. The stimulus duration was 100 ms (Fig. 1c). The measure of interest was the orientation threshold in degrees.

Contrast sensitivity (Con): A two-interval forced-choice test was used (Lahav et al., 2011). Participants viewed subsequent red and green circles at the center of the screen, each with a diameter of 3 arcdeg (Fig. 1d). They were asked to indicate in which of the two circles a Gabor patch was presented. The Gabor patch had a mean luminance of 40 cd/m^2, a spatial frequency of 4 cycles/arcdeg, an envelope sigma of 0.30 arcdeg, and a presentation duration of 100 ms. Dithering techniques were applied to virtually increase the limited gray level resolution of the monitors. The measure of interest was the contrast sensitivity threshold.

Motion direction sensitivity (MotDir): Participants were administered a test similar to the one used by Pilz et al. (2017) to assess motion direction sensitivity. In this test, participants were required to discriminate between rightward or leftward motion in a random dot pattern. The random dot pattern consisted of a certain proportion of dots moving coherently either to the right or left, while the remaining dots moved in random directions (Fig. 1e). The stimulus consisted of 50 white dots (dot size: 5 arcmin) moving at a speed of 5.6 arcdeg per second within a rectangular area (8.9 × 6.7 arcdeg). The stimulus was presented for 5 s. Participants were instructed to indicate the perceived motion direction of the coherently moving dots. The measure of interest was the ratio of coherently moving dots.

Simple reaction time (RT): The test was based on the Hick-paradigm (Hick, 1952). Participants completed 80 trials in which they had to press a push button immediately after the presentation of a white square (size: 3 arcdeg, duration of presentation: 500 ms) on a black background. The inter-trial interval varied randomly between 1.5 and 2 s to prevent anticipation. The measure of interest was the mean reaction time.

Visual search (VSrch): The visual search test was based on the work by Theeuwes and Kooi (1994). Participants searched for a green horizontal line within an array of distractors (green vertical and red horizontal lines; Fig. 1f). Conditions with four, nine or 16 lines (length: 1600 arcsec, width: 450 arcsec) were presented in a random order for a max. duration of 10 s. A total of 120 trials were administered (i.e., 40 trials per condition). Participants indicated whether or not the display contained a green horizontal line, which was present in 50 % of the trials. Average reaction times for correctly answered trials (RTc) and the percentage correct (PC) were recorded and combined into an inverse efficiency score (i.e., RTc / PC; Vandierendonck, 2018), which was then averaged across the three conditions.

Simon effect (Simon): The Simon test was based on the one used by

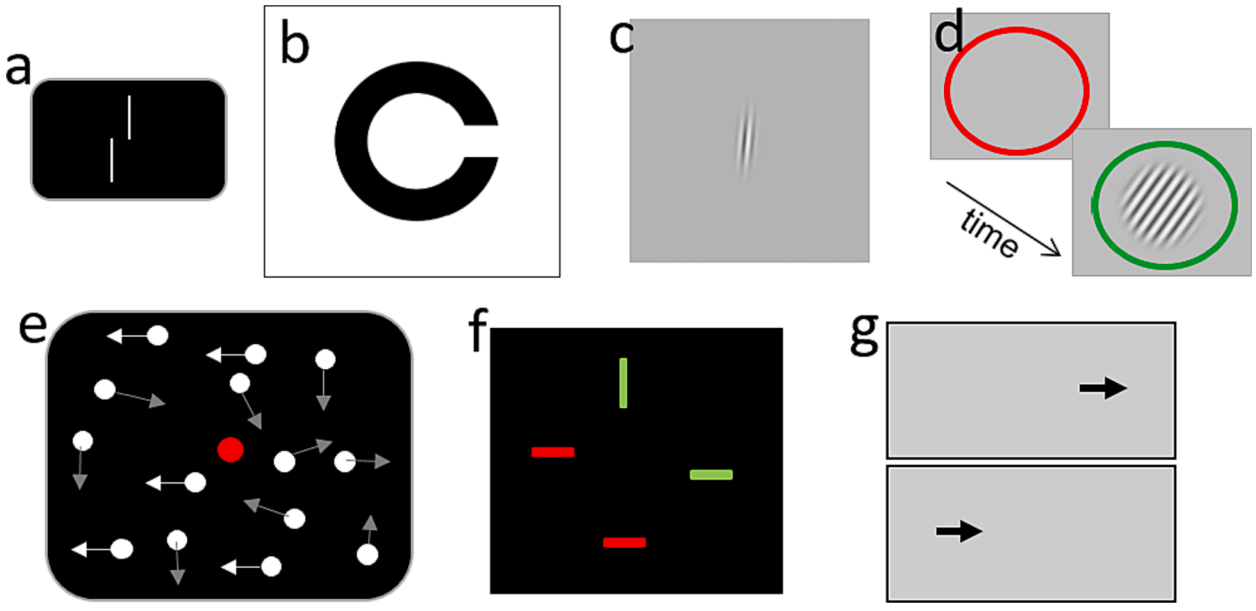

Fig. 1. Illustrations of the tests used except simple reaction time: (a) vernier offset discrimination; (b) Freiburg visual acuity; (c) orientation discrimination; (d) contrast sensitivity; (e) motion direction sensitivity (only a few dots are shown, where arrows are used to depict the test but were not present in the real task); (f) visual search with four lines; and (g) congruent (upper panel) and incongruent (lower panel) Simon test.

Castel et al. (2007). Arrows (length: 0.60 arcdeg) were presented for 100 ms at three different locations on the screen (left, center or right). In congruent trials, the direction of the arrow matched its location (e.g., right-pointing arrow presented on the right side of the screen), while in incongruent trials the direction of the arrow did not match its location (e.g., right-pointing arrow presented on the left side of the screen; Fig. 1g). A control condition with center-presented arrows was also included but not used in the data analysis for this study. There were 40 trials for each condition, presented in random order. Participants were instructed to report the direction of the arrow. The Simon effect (Simon) was calculated as the difference in reaction time between correctly answered incongruent and congruent trials, divided by the average response time.

2.4. Data analysis

For each visual test, we extracted one score for each participant at each visit. Analysis were performed on two datasets: dataset A, which included the scores of the 58 older adults who participated in the firsts two visits, and dataset B, which included the scores of the 37 older adults who participated in all three visits. Analysis was performed in Python.

2.4.1. Preprocessing of visual tests scores

Most scores distributions violated the normality assumption as assessed by the Shapiro-Wilk test, mainly because of skewness (Figures S1, S2 and Tables S2, S3 in the Supplementary Material). Therefore, we wanted to approximate a normal distribution while removing the outliers. To do so, for each visual test, the two scores repetitions of dataset A and the three scores repetitions of dataset B were pooled. The following pre-processing steps were then performed on these pooled scores: (1) we computed modified *z*-scores (which are based on the median and median absolute deviation) and removed outliers according to a 3.5 *SD* criterion (Iglewicz & Hoaglin, 1993) (2) we used the *Yeo-Johnson* power transformation (using the PowerTransformer function from sklearn.preprocessing python package; Pedregosa et al., 2011) and optimized its λ exponent to maximize normality according to the Shapiro-Wilk test (3) we included the previously removed outliers and transformed the variables using the *Yeo-Johnson* transformation with the optimized λ parameter (4) we repeated the outlier removal step as in step 1, and (5) for all visual tests except VA, we flipped the sign of the scores to indicate better performance with higher scores (see Figures S3 and S4 for the resulting scores distributions and Tables S2 and S3 for the pre-processing parameters and results in the Supplementary Material).

Subsequently, we removed participants who displayed exceptional performance instability or stability across visits (in contrast to outliers identified in step 4, indicating exceptionally poor or good test scores). This was achieved by repeating an outlier removal step on the differences between scores from different visits. Specifically, score differences were computed as visit2-visit1 for dataset A, and as visit3-visit1, visit2-visit1 and visit3-visit2 for dataset B. For each dataset, participants were excluded if at least one of the score differences was identified as an outlier (see Tables S2 and S3 in the Supplementary Material).

After removing the outliers, the dataset A had 7.3 % missing scores, and the dataset B had 7.8 % missing scores (Tables S4 and S5 in the Supplementary Material). Participants with missing scores were not excluded, and no data imputation was performed. Instead, pairwise deletion was applied for computing correlations.

2.4.2. Long-term performance stability

To visualize the stability of individual performances in each visual test across visits, we plotted the participants' scores of two visits as scatter plots. We computed Pearson's, Spearman's and intraclass correlations (ICCs) to quantify the long-term performance stability across two visits. Pearson's and Spearman's correlations are often used to evaluate test–retest reliability. However, ICCs are in general conceptually more appropriate for measuring test–retest reliability although in our specific case, where we use ICC(C,1), the measure is very similar to Pearson's correlation. ICC(C,1) estimates the ratio of between-individuals variance to the total variance. The total variance comprises the between-individuals variance, measurement error, and within-individual variance. For example, an ICC(C,1) of 0.5 indicates that the sum of the measurement error and the within-individual variance is as high as the between-individual variance (Liljequist, Elfving, & Roaldsen, 2019; Shrout & Fleiss, 1979).

Although Pearson's correlations and ICCs are conceptually different, in the specific case of ICC(C,1) the values are similar. In fact, we can derive ICC(C,1) from Pearson's formula if, instead of estimating the variances of the two variables separately, the same variance can be assumed for both variables and therefore estimated over the pooled data, as follows:

$$ICC(3,1) = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)\cdot sx\cdot sy}, \text{ where } sx = sy = \sqrt{\frac{\sum(x-\bar{x})^2+\sum(y-\bar{y})^2}{2n-2}}.$$

2.4.3. Between-tests correlations

Pearson and distance correlations were computed to quantify the relationship between the visual tests scores in one visit. Pearson's correlations measure the linear relationship between the scores of two tests, whereas distance correlations also detect non-linear and higher-dimensional correlations (Székely, Rizzo, & Bakirov, 2007).

To examine if the magnitudes of the correlation coefficients changed with time, we compared the Pearson's correlation coefficients across two visits. For this analysis, we used Fisher's r to *Z* transformation: $Z = \frac{1}{2}*(\ln(1+r) - \ln(1-r))$ and compared the *z* test statistics between the two visits as:

$$z = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}}$$

where Z_1 and Z_2 are two *Z* transformed correlations (one for each visit) and n_1 and n_2 are the sample sizes for the same test at the two visits, respectively. We used a two-sided alternative hypothesis with alpha of 0.05.

3. Results

3.1. Long-term performance stability

Fig. 2 shows the scatter plots for each visual test, where each point corresponds to a participant score at visit 1 vs. visit 2 (i.e., four years apart, upper panel), at visit 2 vs. visit 3 (i.e., three years apart, middle panel), and at the visit 1 vs. visit 3 (i.e., seven years apart, lower panel). The x and y axes have the same scale, so points along the diagonal indicate that participants had the same score at the two visits, while points above the diagonal indicate a score improvement over time. Associated ICCs, Pearson's and Spearman's coefficients (i.e., effect sizes) and p-values are reported in Table 2. Dotted diagonals in the scatterplots indicate non-significant ICCs.

Overall, the three comparisons between visits show a consistent pattern. Visual inspection reveals that for most tests, participants who perform better than others in one visit also tend to perform better in the other visit(s). This stability of individual differences between visits was confirmed by many significant long-term test–retest correlations. In general, there was high agreement between Pearson's, Spearman's and intraclass correlations.

To further interpret the strength of long-term test–retest correlations, we examined the effect sizes. Pearson's and Spearman-s correlation coefficients of 0.1, 0.3 and 0.5 are interpreted as small, medium and large following Cohen (1988). Typical interpretation of the ICC is more strict, with ICCs > 0.90 considered as excellent, between 0.75 and 0.90 as good, between 0.50 and 0.75 as moderate and ICCs < 0.50 as poor (Koo & Li, 2016).

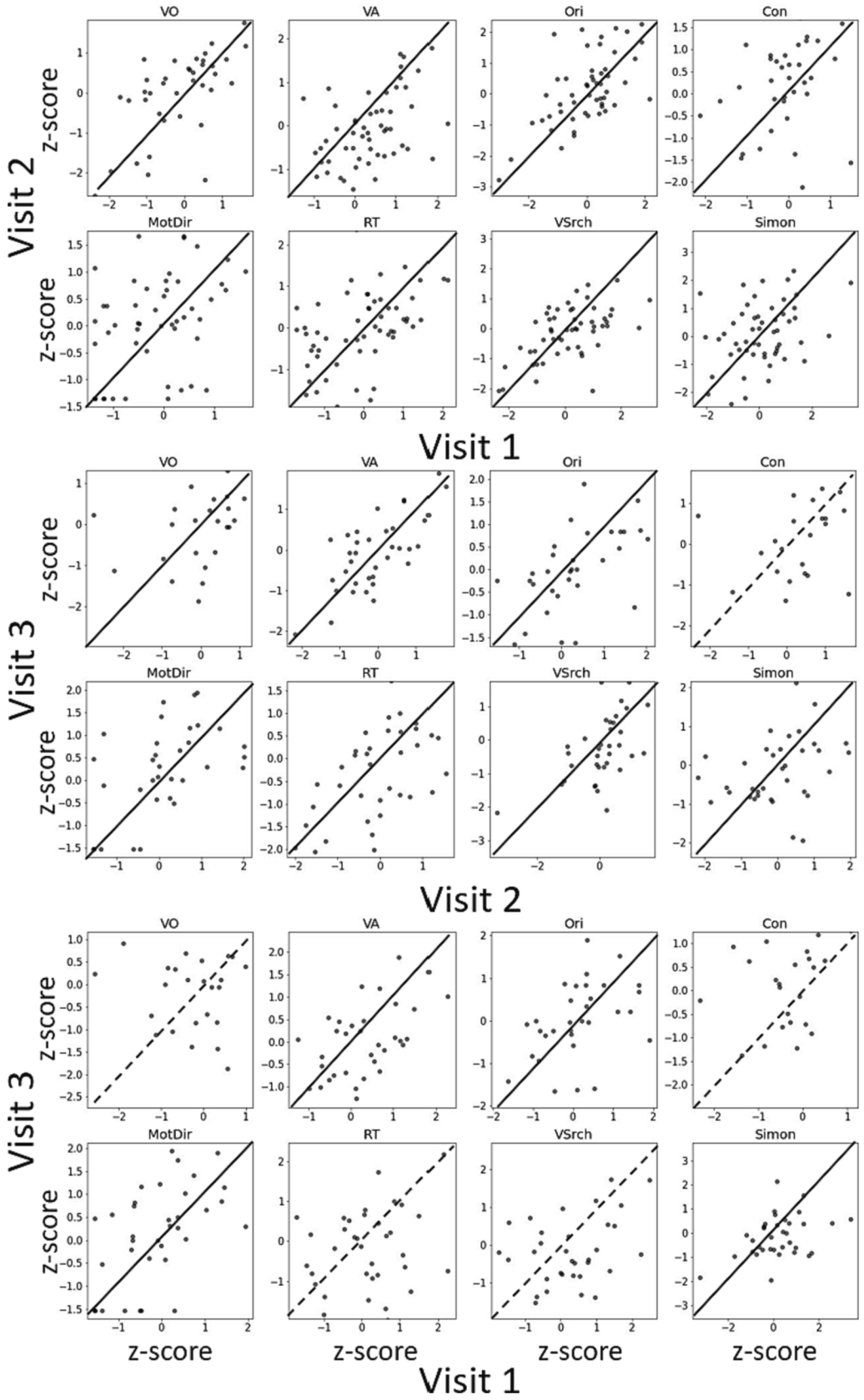

Fig. 2. Scatter plots for all tests for all pairwise comparisons of visits, with the earlier visit on x and the later visit on y. Points along the diagonal indicate that participants achieved the same score in the two visits; points above the diagonal, for example, indicate performance improvement over time. Dotted diagonals are used to indicate non-significant ICCs.

Table 2
Pearson, Spearman, and Intraclass correlation of type (C,1) for visit 1 vs. visit 2, visit 2 vs. visit 3 and visit 1 vs. visit 3.

	visit 1 vs. visit 2 (n = 58)			visit 2 vs. visit 3 (n = 37)			visit 1 vs. visit 3 (n = 37)		
	Pearson r, p	*Spearman r, p*	*ICC31, p*	*Pearson r, p*	*Spearman r, p*	*ICC31, p*	*Pearson r, p*	*Spearman r, p*	*ICC31, p*
VO	0.621, 3.2e-5	0.603, 6.2e-5	0.627, 1.3e-5	0.325, 0.11	0.394, 0.05	0.325, 0.05	−0.110, 0.58	−0.080, 0.70	−0.110, 0.711
VA	0.536, 1.7e-5	0.524, 2.9e-5	0.527, 1.3e-5	0.716, 9.4e-7	0.614, 6.8e-5	0.715, 3.2e-7	0.570, 2.3e-5	0.520, 9.6e-4	0.514, 5.7e-4
Ori	0.613, 1.0e-6	0.540, 3.6e-5	0.624, 3.8e-7	0.541, 1.4e-3	0.615, 1.8e-4	0.541, 5.7e-4	0.586, 2.7e-5	0.556, 6.4e-4	0.582, 1.2e-4
Con	0.391, 0.02	0.415, 9.7e-3	0.368, 0.01	0.260, 0.23	0.375, 0.08	0.260, 0.11	0.215, 0.30	0.232, 0.27	0.209, 0.15
MotDir	0.478, 2.5e-4	0.503, 1.1e-4	0.457, 2.8e-4	0.595, 1.6e-4	0.617, 7.8e-5	0.595, 6.6e-5	0.592, 1.1e-5	0.594, 1.1e-4	0.564, 1.4e-4
RT	0.428, 7.9e-4	0.451, 3.8e-4	0.409, 7.8e-4	0.573, 2.1e-4	0.544, 5.1e-4	0.570, 9.7e-5	0.253, 0.13	0.197, 0.24	0.227, 0.09
VSrch	0.559, 1.1e-5	0.502, 1.1e-4	0.553, 7.0e-6	0.554, 5.5e-4	0.587, 2.1e-4	0.552, 2.5e-4	0.261, 0.14	0.212, 0.23	0.259, 0.07
Simon	0.269, 0.05	0.231, 0.09	0.250, 0.03	0.304, 0.07	0.363, 0.03	0.300, 0.04	0.396, 0.01	0.352, 0.04	0.375, 0.01

Listwise deletion was used. Effect sizes (*r*, ICC31) and associated p-values (*p*) of Pearson, Spearman, and intraclass correlations are reported.

Here, the values of the three correlation coefficient types were very similar. Hence, we interpret them collectively and use 0.5 as the criterion for identifying meaningful effect sizes. For test–retest over four years (i.e., visit 1 vs. visit 2), four out of eight effect sizes (Pearson's, Spearman's and intraclass correlations combined) were larger than 0.50, and none was smaller than 0.23. Over three years (i.e., visit 2 vs. visit 3), five out of eight effect sizes were larger than 0.50, and none was smaller than 0.26. Even over seven years (i.e., visit 1 vs. visit 3), three out of eight effect sizes were larger than 0.5. Only two tests did not show effect sizes larger than 0.5 in any of the comparisons between visits: Simon and contrast test.

3.2. Between-tests correlations

Between-tests Pearson's correlation coefficients *r* are shown in the left panel of Fig. 3. The portion of the correlation matrix below the diagonal (i.e., lower part) represents correlations for visit 1, and the portion of the correlation matrix above the diagonal (i.e., upper part) represents correlations for visit 2. The correlation coefficients *r* for visit 3 are shown in the lower part of the correlation matrix of Figure S6 in the Supplementary Materials. Overall, correlations were weak and mostly non-significant for each visit. The 25th, 50th and 75th percentiles of *r* were as follows: 0.10, 0.20 and 0.34 for visit 1; 0.04, 0.18 and 0.30 for visit 2; and −0.06, 0.07 and 0.26 for visit 3. Principal component analysis (PCA) was also conducted to detect multivariable relationships, and the percentage of variance explained by the first eigenvalue was extracted. This value represents the amount of variance that could be explained if only the largest factor is retained. The results indicated that the 1st eigenvalue explained only 33 % of the variance in visit 1, 30 % in visit 2, and 26 % in visit 3. Overall, distance correlations revealed the same pattern (Figure S5 for visit 1 and visit 2, upper part of Figure S6 for visit 3 in the Supplementary Materials). These results suggest no evidence for a visual common factor in any of the visits.

Although in general the correlations were weak, two out of the 28 computed between-tests correlations showed moderate strength (i.e., $r > 0.3$) in each visit: VA vs. VO ($r = 0.41$ for visit 1, $r = 0.37$ for visit 2, and $r = 0.33$ for visit 3), and Ori vs. VO ($r = 0.54$ for visit 1, $r = 0.31$ for visit 2, and $r = 0.50$ for visit 3).

Fisher's *r*-to *Z* transformation was used to statistically compare the correlation coefficients between visit 1 and visit 2, which basically yields z-transformed differences of the corresponding correlation coefficients. Only very few of these z-values, displayed in Fig. 3, indicated significant differences between the correlation coefficients of the two visits: one coefficient was larger in visit 2, while two were smaller. Please note that we did not compare the correlation matrices with visit 3, as visit 3 included only a subset of participants.

4. Discussion

In vision, there are mainly weak correlations between the performance levels of most visual paradigms (Mollon, Bosten, Peterzell, & Webster, 2017; Peterzell, 2016; Tulver, 2019). The first possible explanation is that tests have a low test–retest reliability (Wang & De Boeck, 2022). However, previous studies found high short-term test–retest reliability for many paradigms when the paradigms were tested twice, usually on the same or next day (Bargary et al., 2017; Bosten & Mollon,

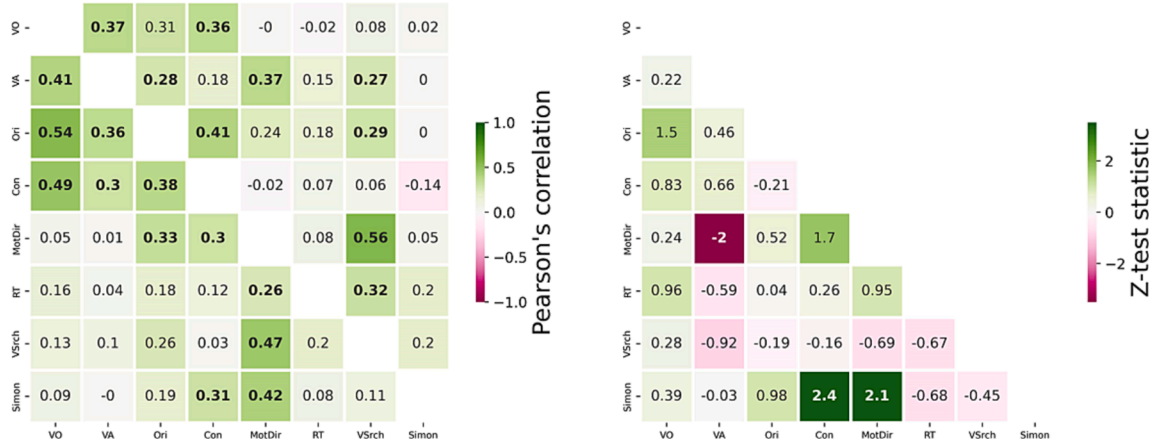

Fig. 3. Left panel: Between-tests Pearson correlation coefficients for visit 1 (lower part) and visit 2 (upper part). Bold numbers indicate significant results ($p < 0.05$, we did not correct for multiple comparisons). The color scale ranging from pink to green represents effect sizes from $r = -1$ to $r = 1$ (white corresponds to $r = 0$). Please note that pairwise deletion was used to compute correlations, thus, the level of significance slightly differs because of the variations in sample size. Right panel: *z* test statistics used to compare the correlation coefficients between visit 1 and visit 2. Bold indicates a significant *z* ($p < 0.05$). The color scale ranging from pink to green reflects *z* values from $z = -3$ to $z = 3$ (white corresponds to $z = 0$). Positive *z* values indicate that the correlation coefficient was larger in visit 1 compared to visit 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2010; Brascamp et al., 2019; Cappe et al., 2014; Cretenoud et al., 2019; Grzeczkowski et al., 2017; Shaqiri et al., 2019; Tulver et al., 2019). The second possible explanation is that between-participants variance is low, i.e., participants perform more or less on the same level, and the residual noise leads to these low correlations. However, between-participants variance is usually rather large for most tests. A third explanation is that the weak correlations are due to strong random fluctuations in the performance levels within each participant. In this scenario, performance is stable when tested on the same day or subsequent days, i.e., short-term test–retest reliability is high, but performance strongly fluctuates across months and years for each participant and each test. Indeed, there is evidence for such long-term fluctuations (Wexler et al., 2015). However, our study shows that there is little evidence for this scenario for our tests, since performance is stable across years for each test and each participant. It seems that the low correlations are caused by large but stable individual differences.

For a few of our tests, the situation appears to be different though. For the Simon and the contrast sensitivity test, long-term test–retest correlations were below 0.5 in all comparisons between visits. However, for both tests, we think that the long-term performance stability was low not necessarily because of the individual performance changes, but because the tests had low test–retest reliability in the first place. Indeed, the low reliability values for the Simon test might be attributed to the reliability paradox, that is, to the reduction of between-participants variance that is caused by how the difference score is derived from the incongruent and congruent conditions (Goodhew & Edwards, 2019; Hedge, Powell, & Sumner, 2018). Regarding contrast sensitivity, it is possible that the 2-interval forced choice task was confusing, particularly for healthy older adults, because it demands more cognitive resources unlike the simpler binary tasks (Yeshurun, Carrasco, & Maloney, 2008). All other tests showed long-term test–retest correlations greater than 0.5 in at least one of the comparisons between visits. Not surprisingly, these values were generally lower than the short-term test–retest reliabilities reported in previous studies, which average around 0.7 (Cappe et al., 2014; Cretenoud et al., 2019; Grzeczkowski et al., 2017). This is because short-term test–retest mainly reflects the methodological measurement error, while long-term test–retest correlations also includes within-individual performance changes. Remarkably, these within-individual performance changes were small enough to detect consistent individual differences throughout seven years. It is important to note that the long-term performance stability reported in our study represent the minimum expected stability for these tests due to several factors: a small sample size, testing healthy older adults who may experience more individual age-related decline than young adults, and modifications to the setup after the first visit due to technical issues.

There are few studies that have also reported stable long-term performance stability for visual tests other than those used in the current study. Kosovicheva and Whitney (2017) found stable individual performance in an object localization task over six months. Wexler et al. (2015) found that, even though there were significant long-term fluctuations in a few individuals, performance in response to ambiguous stimuli was stable for most individuals one year after the initial testing. In a follow-up experiment, participants were tested daily for three months, and the results showed that biases in ambiguous stimuli evolve over time following a random walk pattern. In other words, responses on trials closer together in time tend to be closer than responses farther apart. These findings are in line with the current study, where performances were found to be more stable over three and four years of testing compared to seven years. Future research should consider conducting a study similar to the one conducted by Wexler et al. (2015) in order to investigate the time-series changes in performance across a comprehensive battery of visual tests.

Obviously, the between-tests correlations cannot be higher than the test–retest correlations. However, the between-tests correlations in the current study averaged around 0.19, which is weak not only in absolute terms but also in relation to the test-retests. The latter averaged around 0.70 for short-term test–retest, as reported in previous studies, and around 0.45 for long-term test–retest in our current study. This large relative difference provides counter-evidence for a visual common factor, which would manifest itself in correlation coefficients much closer to the test–retest coefficients. It's worth noting that two moderate correlations were found in the current study between tests that require spatial vision, namely between VA and VO, and between Ori and VO. Interestingly, the correlation between Ori and VO was also moderate in young adults ($r = 0.32$; Shaqiri et al., 2019), whereas no correlation was found between VA and VO in young adults ($r = 0.05$ in Shaqiri et al., 2019; $r = 0.03$ in Cappe et al., 2014).

A limitation of this study is the lack of testing short-term test–retest reliability, which would have allowed for disentangling measurement error from within-individual variance by comparing the short-term to the long-term test–retest correlations.

It is important to note that some visual tests commonly used in vision research showed poor short-term test–retest reliabilities (Chamberlain, Van der Hallen, Huygelier, Van de Cruys, & Wagemans, 2017; Clark et al., 2022; Milne & Szczerbinski, 2009). Furthermore, in the current study, the Simon and the contrast sensitivity tests had a low long-term test–retest. Therefore, it remains crucial to continue measuring the test–retest reliability of tests, particularly in individual difference research (Hedge et al., 2018; Wang & De Boeck, 2022).

Given the stable individual differences we found across seven years, one might wonder how the results align with the individual age-related decline. The literature suggests that the most prominent age-related visual decline typically occurs after the age of 70 (Arena, Hutchinson, & Shimozaki, 2012; Bennett, Sekuler, & Sekuler, 2007; Brabyn, Schneck, Haegerstrom-Portnoy, & Lott, 2001). Indeed, we tested participants in their 60 s and observed only modest performance changes (Fig. 2 and Tables S4 and S5 in the Supplementary Material). A larger decline does not necessarily impact long-term test–retest correlations, but it does if the decline is individually different. This could be investigated in future research by testing healthy adults older than 70 years longitudinally. By applying a battery of visual tests one can also study whether the (individual) decline is homogenous across different visual abilities.

To conclude, we found weak correlations between visual abilities at different points in time throughout seven years despite long-term performance stability. Thus, the performance in one test can predict the performance in the same test even after as long as seven years but is uninformative of the performance in a different test.

Funding

This work was funded by the National Centre of Competence in Research (NCCR) Synapsy financed by the Swiss National Science Foundation under grant 51NF40-185897. The funding source had no role in study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

CRediT authorship contribution statement

Simona Garobbio: Formal analysis, Conceptualization, Writing – original draft. **Marina Kunchulia:** Data curation, Writing – review & editing. **Michael H. Herzog:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We would like to thank Marc Repnow for technical help in the study.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.visres.2023.108355.

References

Arena, A., Hutchinson, C. V., & Shimozaki, S. S. (2012). The effects of age on the spatial and temporal integration of global motion. *Vision Research, 58*, 27–32. https://doi.org/10.1016/j.visres.2012.02.004

Bach, M. (1996). The Freiburg Visual Acuity Test-automatic measurement of visual acuity. *Optometry and Vision Science, 73*(1), 49–53. https://doi.org/10.1097/00006324-199601000-00008

Bargary, G., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Hogg, R. E., & Mollon, J. D. (2017). Individual differences in human eye movements: An oculomotor signature? *Vision Research, 141*, 157–169. https://doi.org/10.1016/j.visres.2017.03.001

P.J. Bennett R. Sekuler A.B. Sekuler The effects of aging on motion detection and direction identification Vision Research 47 6 2007 799 809 https://doi.org/https://doi.org/10.1016/j.visres.2007.01.001.

Bocheva, N., Angelova, D., & Stefanova, M. (2013). Age-related changes in fine motion direction discriminations. *Experimental Brain Research, 228*(3), 257–278. https://doi.org/10.1007/s00221-013-3559-4

Bosten, J. M., & Mollon, J. D. (2010). Is there a general trait of susceptibility to simultaneous contrast? *Vision Research, 50*(17), 1656–1664. https://doi.org/10.1016/j.visres.2010.05.012

Bosten, J. M., Mollon, J. D., Peterzell, D. H., & Webster, M. A. (2017). Individual differences as a window into the structure and function of the visual system. *Vision Research, 141*, 1–3. https://doi.org/10.1016/j.visres.2017.11.003

Brabyn, J., Schneck, M., Haegerstrom-Portnoy, G., & Lott, L. (2001). The Smith-Kettlewell Institute (SKI) longitudinal study of vision function and its impact among the elderly: An overview. *Optometry and Vision Science, 78*(5), 264–269. https://doi.org/10.1097/00006324-200105000-00008

Brascamp, J. W., Qian, C. S., Hambrick, D. Z., & Becker, M. W. (2019). Individual differences point to two separate processes involved in the resolution of binocular rivalry. *Journal of Vision, 19*(12), 1–17. https://doi.org/10.1167/19.12.15

Cappe, Ć., Clarke, A., Mohr, C., & Herzog, M. H. (2014). Is there a common factor for vision? *Journal of Vision, 14*(8), 1–11. https://doi.org/10.1167/14.8.4

Castel, A. D., Balota, D. A., Hutchison, K. A., Logan, J. M., & Yap, M. J. (2007). Spatial attention and response control in healthy younger and older adults and individuals with Alzheimer's disease: Evidence for disproportionate selection impairments in the Simon task. *Neuropsychology, 21*(2), 170–182. https://doi.org/10.1037/0894-4105.21.2.170

Chamberlain, R., Van der Hallen, R., Huygelier, H., Van de Cruys, S., & Wagemans, J. (2017). Local-global processing bias is not a unitary individual difference in visual processing. *Vision Research, 141*, 247–257. https://doi.org/10.1016/j.visres.2017.01.008

Clark, K., Birch-hurst, K., Pennington, C. R., Petrie, A. C. P., Lee, J. T., & Hedge, C. (2022). Test-retest reliability for common tasks in vision science. *Journal of Vision, 22*(8), 1–18. https://doi.org/doi: 10.1167/jov.22.8.18.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (Second)*. Academic Press.

Cretenoud, A. F., Karimpur, H., Grzeczkowski, L., Francis, G., Hamburger, K., & Herzog, M. H. (2019). Factors underlying visual illusions are illusion-specific but not feature-specific. *Journal of Vision, 19*(14), 1–21. https://doi.org/10.1167/19.14.12

Delahunt, P. B., Hardy, J. L., & Werner, J. S. (2008). The effect of senescence on orientation discrimination and mechanism tuning. *Journal of Vision, 8*(3), 1–9. https://doi.org/10.1167/8.3.5

Frenzel, H., Bohlender, J., Pinsker, K., Wohlleben, B., Tank, J., Lechner, S. G., … Lewin, G. R. (2012). A genetic basis for mechanosensory traits in humans. *PLoS Biology, 10*(5). https://doi.org/10.1371/journal.pbio.1001318

Garobbio, S., Pilz, K. S., Kunchulia, M., & Herzog, M. H. (2022). No Common Factor Underlying Decline of Visual Abilities in Mild Cognitive Impairment. *Experimental Aging Research, 1–18*. https://doi.org/10.1080/0361073X.2022.2094660

Goodhew, S. C., & Edwards, M. (2019). Translating experimental paradigms into individual-differences research: Contributions, challenges, and practical recommendations. *Consciousness and Cognition, 69*(January), 14–25. https://doi.org/10.1016/j.concog.2019.01.008

Grzeczkowski, L., Clarke, A. M., Francis, G., Mast, F. W., & Herzog, M. H. (2017). About individual differences in vision. *Vision Research, 141*, 282–292. https://doi.org/10.1016/j.visres.2016.10.006

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

Hick, W. E. (1952). On the Rate of Gain of Information. *Quarterly Journal of Experimental Psychology, 4*(1), 11–26. https://doi.org/10.1080/17470215208416600

Iglewicz, B., & Hoaglin, D. C. (1993). Volume 16: How to detect and handle outliers, The ASQC basic references in quality control: Statistical techniques. *Edward F. Mykytka, 16*. https://doi.org/10.2307/1269377

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

A. Kosovicheva D. Whitney Stable individual signatures in object localization Current Biology 27 14 2017 R700 R1 10.1016/j.cub.2017.06.001.

Lahav, K., Levkovitch-Verbin, H., Belkin, M., Glovinsky, Y., & Polat, U. (2011). Reduced mesopic and photopic foveal contrast sensitivity in glaucoma. *Archives of Ophthalmology (Chicago, Ill. : 1960)*, *129*(1), 16–22. https://doi.org/10.1001/archophthalmol.2010.332.

Liljequist, D., Elfving, B., & Roaldsen, K. S. (2019). Intraclass correlation – A discussion and demonstration of basic features. In. *PLoS ONE, 14*. https://doi.org/10.1371/journal.pone.0219854

Milne, E., & Szczerbinski, M. (2009). Global and local perceptual style, field-independence, and central coherence: An attempt at concept validation. *Advances in Cognitive Psychology, 5*(1), 1–26. https://doi.org/10.2478/v10053-008-0062-8

Mollon, J. D., Bosten, J. M., Peterzell, D. H., & Webster, M. A. (2017). Individual differences in visual science: What can be learned and what is good experimental practice? *Vision Research, 141*(November), 4–15. https://doi.org/10.1016/j.visres.2017.11.001

Owsley, C., & McGwin, G. (2010). Vision and driving. *Vision Research, 50*(23), 2348–2361. https://doi.org/10.1016/j.visres.2010.05.021

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Retrieved from *Journal of Machine Learning Research, 12*(85), 2825–2830 http://jmlr.org/papers/v12/pedregosa11a.html.

Peterzell, D. H. (2016). Discovering sensory processes using individual differences: A review and factor analytic manifesto. *Human Vision and Electronic Imaging 2016. HVEI, 2016*, 110–120. https://doi.org/10.2352/ISSN.2470-1173.2016.16HVEI-112

Pilz, K. S., Miller, L., & Agnew, H. C. (2017). Motion coherence and direction discrimination in healthy aging. *Journal of Vision, 17*(1), 1–12. https://doi.org/10.1167/17.1.31

Plomp, G., Kunchulia, M., & Herzog, M. H. (2012). Age-related changes in visually evoked electrical brain activity. *Human Brain Mapping, 33*(5), 1124–1136. https://doi.org/10.1002/hbm.21273

Scialfa, C. T., Esau, S. P., & Joffe, K. M. (1998). Age, target-distractor similarity, and visual search. *Experimental Aging Research, 24*(4), 337–358. https://doi.org/10.1080/036107398244184

Shaqiri, A., Pilz, K. S., Cretenoud, A. F., Neumann, K., Clarke, A., Kunchulia, M., & Herzog, M. H. (2019). No Evidence for a Common Factor Underlying Visual Abilities in Healthy Older People. *Developmental Psychology, 55*(8), 1775–1787. https://doi.org/10.1037/dev0000740

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

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics, 35*(6), 2769–2794. https://doi.org/10.1214/009053607000000505

Taylor, M., & Creelman, C. D. (1967). PEST: Efficient estimates on probability functions. *The Journal of the Acoustical Society of America, 41*(4a), 782–787. https://doi.org/10.1121/1.1910407

Theeuwes, J., & Kooi, F. L. (1994). Parallel search for a conjunction of contrast polarity and shape. *Vision Research, 34*(22), 3013–3016. https://doi.org/10.1016/0042-6989(94)90274-7

Tibber, M. S., Guedes, A., & Shepherd, A. J. (2006). Orientation discrimination and contrast detection thresholds in migraine for cardinal and oblique angles. *Investigative Ophthalmology and Visual Science, 47*(12), 5599–5604. https://doi.org/10.1167/iovs.06-0640

Tulver, K. (2019). The factorial structure of individual differences in visual perception. *Consciousness and Cognition, 73*(June), Article 102762. https://doi.org/10.1016/j.concog.2019.102762

Tulver, K., Aru, J., Rutiku, R., & Bachmann, T. (2019). Individual differences in the effects of priors on perception: A multi-paradigm approach. *Cognition, 187*(March), 167–177. https://doi.org/10.1016/j.cognition.2019.03.008

Vandierendonck, A. (2018). Further Tests of the Utility of Integrated Speed-Accuracy Measures in Task Switching. *Journal of Cognition, 1*(1), 1–16. https://doi.org/10.5334/joc.6

Verhallen, R. J., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Bargary, G., & Mollon, J. D. (2017). General and specific factors in the processing of faces. *Vision Research, 141*, 217–227. https://doi.org/10.1016/j.visres.2016.12.014

Wang, S., & De Boeck, P. (2022). Understanding the role of subpopulations and reliability in between-group studies. *Behavior Research Methods, 54*(5), 2162–2177. https://doi.org/10.3758/s13428-021-01700-8

Wechsler, D. (2003). *Wechsler intelligence scale for children* (4th ed). San Antonio, TX: The Psychological Corporation.

Wexler, M., Duyck, M., & Mamassian, P. (2015). Persistent states in vision break universality and time invariance. *Proceedings of the National Academy of Sciences of the United States of America, 112*(48), 14990–14995. https://doi.org/10.1073/pnas.1508847112.

Yeshurun, Y., Carrasco, M., & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research, 48*(17), 1837–1851. https://doi.org/10.1016/j.visres.2008.05.008