Applied Psychology Research 2025, 4(2), 2345.
doi: https://doi.org/10.59400/apr2345

Article

An evaluation of the required number of training sessions of neuropsychological assessments on portable mobile devices

Jim Jansen[1], Aurora JAE van de Loo[1], Johan Garssen[1,2], Andrew Scholey[3,4,5], Brian Tiplady[6], Joris C. Verster[1,5,7,*]

[1] Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, 3584 CG Utrecht, The Netherlands
[2] Danone Global Research & Innovation Center, Uppsalalaan 12, 3584 CT Utrecht, The Netherlands
[3] School of Psychology, Northumbria University, Newcastle upon Tyne NE1 8ST, UK
[4] Nutrition Dietetics and Food, School of Clinical Sciences, Monash University, Melbourne, VIC 3168, Australia
[5] Centre for Mental Health and Brain Sciences, Swinburne University of Technology, Melbourne, VIC 3122, Australia
[6] Department of Anaesthesia, Critical Care & Pain Medicine, University of Edinburgh, Edinburgh EH16 4SA, UK
[7] Cognitive Neurophysiology, Department of Child and Adolescent Psychiatry, Faculty of Medicine, TU Dresden, 01307 Dresden, Germany
*** Corresponding author:** Joris C. Verster, j.c.verster@uu.nl

CITATION

Jansen J, van de Loo AJ, Garssen J, et al. An evaluation of the required number of training sessions of neuropsychological assessments on portable mobile devices. Applied Psychology Research. 2025; 4(2): 2345. https://doi.org/10.59400/apr2345

ARTICLE INFO

Received: 21 December 2024
Revised: 12 June 2025
Accepted: 19 June 2025
Available online: 21 August 2025

Abstract: In some research, it is important to conduct cognitive assessments in an everyday setting. Both tablet PCs and mobile phones have been used in this context. The purpose of this study was to examine whether a mobile test battery yields similar results on a mobile phone (screen size 6 cm diagonal) and a tablet (18 cm). Thirty-nine healthy volunteers (aged 18–30) completed five training sessions and one final "test" session per device. The 18-minute test battery consisted of six tests, measuring attention (Number Pairs Test, NP and Arrow Flankers Test, AF), psychomotor functioning (Arrow reaction time test, AR), working memory (Memory scanning test, MS), paired associate learning (Shape pairs, SP), and comprehension (Serial sevens, SS). Outcome measures were mean reaction time (RT) and the percentage of errors. RT scores over the practice runs indicated that AR and AF required only a single familiarization run, while other tests needed 3–4 runs to achieve stable performance. No difference was seen in practice effects between the platforms. Test scores were similar for the platforms with minimal differences between phone and tablet scores (effect sizes < 0.25). Correlations between phone and tablet scores were in the range 0.53–0.82, except one measure, SP errors, where the correlation was much lower. Taken together, these results indicate that there is generally good agreement between data obtained from phones and tablets with very different screen sizes. Phones with small screens are suitable for assessing cognition in an everyday setting. Training on the tests is recommended to achieve stable performance before the start of experimental sessions.

Keywords: validation; mobile phone; android tablet; neuropsychological function; psychomotor; attention; memory

1. Introduction

Over the past three decades, the use of computer systems to assess cognitive and psychomotor function has become commonplace in clinical investigations. Computer-based testing has many advantages over traditional 'pencil-and-paper' methods, allowing standardized test presentation, better resolution of response times, and simplified data handling (Sternin et al., 2019). On the other hand, it is important

to also consider potential disadvantages of computerized testing, such as privacy and data safety issues, and cultural and age differences in access to, and familiarity with, computerized devices (Bauer et al., 2012).

Neuropsychological testing can be used in a variety of settings. Dedicated psychological testing laboratories allow the greatest control over the environment and the greatest range of testing equipment and methods. This approach generally requires a trained administrator to instruct and monitor participants. The participants are usually required to come to the trial center to complete the tests, which can be time-consuming for both patient and clinical practitioner (Meyers and Brown, 2006). The latter is one of the reasons why an increasing number of so-called naturalistic studies consider 'at home' testing when this is feasible (Verster et al., 2012).

Increasingly, tests are used in clinical settings or other environments, such as workplaces, where equipment needs to fit into a schedule and space that includes many other activities. For such purposes, smaller devices such as tablets and personal digital assistants have many advantages (Tiplady, 1994; Lamond et al., 2005; Vincent et al., 2017). This has led to the development of capabilities to assess participants in an everyday setting, using an ecological momentary assessment (EMA) approach (Stone and Shiffman, 1994; Waters et al., 2012). For this type of assessment, portability and ease of use come even more to the fore, and mobile phones (either "feature" phones or smart phones) are now generally used for this purpose (Thomson et al., 2009; Sliwinski et al., 2018).

The range of screen sizes that are used by these platforms varies considerably, from about 6 cm for feature phones to about 25 cm for the larger tablet PCs. Clearly, not all cognitive tests are suitable for presentation on smaller mobile phone screens, but a range of aspects of cognitive function have been successfully evaluated on these devices, including attention, reaction time, memory, concentration, drug-related cognition, and comprehension (Thomson et al., 2009; Keenan et al., 2014; Jones et al., 2018). The question arises as to the equivalence of similar tests presented on larger and smaller screens, as well as the validity of these portable measures.

The present study is a comparison between a test battery set up on an Android 7-inch diagonal screen (18 cm) tablet and a mobile phone with a keypad and a 2.4-inch (6 cm) screen. Tests were selected from the PenScreen battery (Mobile Cognition Ltd) based on being suitable for both platform sizes, assessing as broad a range of cognitive domains as practicable, and being sensitive to the effects of relevant state changes, such as the acute effects of drugs and alcohol. Although the validity of the tests has been established on particular platforms, no systematic comparison has been carried out to evaluate comparability between different platforms.

The test battery was originally designed for Nokia phones. However, in daily life, Nokia devices have been largely replaced by smartphones and tablets with larger screen sizes. It is therefore important to investigate whether the psychometric tests can be adequately performed on devices with a larger screen size. Of course, it is not possible to evaluate all the different types of smartphones and tablets. Therefore, we chose to compare the Nokia phones with a tablet, having the largest possible range in screen size.

In particular, smaller screens might affect the ability of users to effectively meet task demands. This may be due to relative difficulties in either stimulus processing or response execution. That is, a smaller interface may produce visual stimuli that are more difficult to process visually, or it may be more challenging to execute fine motor movements during task responding. If either or both of these hold true, then one might also expect task performance to be more effortful or require more attentional resources, which could itself negatively affect performance. The 6 cm screen of the Nokia is the smallest that is likely to be used for this type of test, and there is considerable data from the use of this size of phone (Sliwinski et al., 2018). Therefore, it can be assumed that if the tests are unaffected by the size of the screen over this large range of sizes (approximately a ninefold difference in screen area), intermediate-sized devices such as most smartphones are unlikely to perform differently. An additional reason for the choice of device was our intended use in simulated diving studies, where smartphones will not withstand the higher air pressure in hyperbaric chambers.

The primary aim of the study was to determine whether performance on the Nokia phone and tablet differs. A second aim of the study was to evaluate the relative subjective effort used to perform tasks across the platforms. A third aim was to establish how many training sessions are necessary to attain baseline performance levels. It was hypothesized that:

H1. *Performance at both devices would be comparable.*

H2. *Performance on the phone and tablet would be associated with a similar amount of effort.*

H3. *Three to five training sessions would be warranted on both platforms.*

2. Materials and methods

The study used a within-subject crossover design. Healthy volunteers, 18–30 years old, were recruited to participate in this study. A total of 39 students (20 females and 19 males) with a mean (SD) age of 22.1 (2.7) years completed the study procedures. The University of Groningen Psychology Ethics Committee approved the study (approval code: ppo-014-224, approval date: 30 April 2015), and written informed consent was obtained from all participants. The study was conducted in accordance with the guidelines of the Declaration of Helsinki and its latest amendments.

A maximum of 4 subjects were tested on each test day. To rule out a possible impact on performance of alcohol, drugs, medicines, or caffeinated beverages that can influence neuropsychological performance, strict criteria were set for participants. For the same reason, using the Groningen Sleep Quality Scale (Mulder-Hajonides van der Meulen et al., 1981), previous night sleep characteristics were assessed to rule out the impact of inadequate sleep on performance. Subjects were instructed not to consume alcohol for 24 h before testing and not to consume caffeine on the day of testing. Subjects were excluded in case of mental or physical illness, the use of CNS drugs, smoking, current drug use, or a positive breath alcohol test.

Six core tests, described below, were administered. Subjects were randomly selected to start with the tablet or phone. Previous research revealed that practice

effects were seen after four subsequent test sessions (Collie et al., 2003). Therefore, in the current study, five training sessions were completed per device, followed by a final experimental session. The duration of the test day was approximately 5 h, and the duration of the practice procedures took 140 min.

2.1. Administered tests

The six tests (PenScreenSix, Mobile Cognition Ltd., Edinburgh) were conducted on 7-inch touchscreen Tesco Hudl HT7S3 Android 4.2.2 tablets, and Java-enabled Nokia 301 (RM840) Symbian S40 mobile phones. The devices used in this study were provided by Mobile Cognition. Each test takes about 3 min to complete (i.e., 18 min to complete the full test battery). Subjects were instructed to respond as quickly as possible, while avoiding making errors. Speed and accuracy scores (mean reaction time and percentage of errors, respectively) were recorded for each test. The six tests were taken in the same order on both devices. Alternate tests were used in each session: the tests generate new random sequences, and, where relevant, generate matched stimulus sets at random for each presentation. Breaks of 15 and 20 min were scheduled after sessions two and four, respectively. A small break (5 min) was scheduled between the other sessions. Thus, the test battery was practiced 5 times, within a time frame of 140 min. The remaining time of the 5-hour test day was used for intake/screening, a lunch break, and conducting the final 6th test session. Performance on each device was followed by completing the mental effort scale, a visual analog scale (VAS) ranging from 0 ('absolutely no effort') to 10 ('extremely much effort') (Zijlstra and van Doorn, 1985). The effort scale was included to evaluate whether effort on each subsequent test battery was equal. This was deemed important to investigate the true learning effect on the tests and rule out a possible impact of fatigue, which would result in more effort to complete the same tests after repeated administration. Also, it could be hypothesized that different screen sizes of the devices may result in different levels of effort needed to perform tests. For example, a smaller device might be more demanding than a larger one, both in terms of display and response. **Figure 1** gives an overview of the 6 core tests.

2.1.1. Arrow reaction time test (AR)

This was a simple choice reaction task to assess psychomotor functioning (Thomson et al., 2009). An arrow appeared on the screen pointing to the left or right. The subject responded by pressing the left or right key corresponding to the direction of the arrow. Mean reaction time and percentage of errors were recorded.

2.1.2. Number pairs test (NP)

This was a test of attention in the presence of distracting information using the flanker paradigm (Eriksen and Eriksen, 1974). The subject saw an array of five digits on the screen. The task was to decide if the second and fourth digits were the same, and to press the YES button if so, or the NO button if not. The remaining digits (distractors) in the sequence could be different from the target digits (neutral) or the same as the target digits (active). Mean reaction time and percentage of errors were recorded.

Test	Example	Instruction
Arrow Reaction Time (AR)	→	Press the RIGHT (yes) or LEFT (no) button corresponding to the direction of the arrow.
Number Pairs (NP)	26461 27761	Press the RIGHT (yes) or LEFT (no) button to indicate if the 2nd and 4th digit are the same or not.
Arrow Flanker (AF)	→→→→→ □□←□□ ××→××	Press the RIGHT (yes) or LEFT (no) button corresponding to the direction of the central arrow. When flanker symbols are crosses, no response should be made.
Memory Scanning (MS)	74283 7	Upper display: Number set to be remembered; Lower display: Stimulus digit Press RIGHT (yes) or LEFT (no) to indicate if the shown digit is part of the previously shown set of five digits.
Shape Pairs (SP)	□ △ △	Upper display: Pair of shapes to be remembered; Lower display: Stimulus shape Press RIGHT or LEFT to indicate the position of a shape within a previous shown pair of shapes. Up to 8 sets of pairs have to be memorized.
Serial Sevens (SS)	834 – 827 – 818 – 811 – 806 – 799 ---	Press RIGHT (yes) or LEFT (no) to indicate whether the shown number is 7 less than previous number or not.

Figure 1. Illustrations of the core test battery.
Note: In the middle column, each item surrounded by a box represents a single screen display.

2.1.3. Arrow flanker test (AF)

This is a test of attention in the presence of distracting information, but using an array of five symbols, rather than digits (Eriksen and Eriksen, 1974). The central symbol was always an arrow, pointing to the right or left. If the arrow pointed to the left, the subject had to press the left arrow button, and if it pointed to the right, the right arrow key had to be pressed. The four flanker symbols could also be arrows, pointing either in the same or the opposite direction as the target; neutral (squares) or suppressors (crosses). The task is to press a left or right button corresponding to the direction of the target as quickly as possible, unless the flankers are crossed, in which case no response should be made. Mean reaction time and percentage of errors were recorded.

2.1.4. Memory scanning test (MS)

This was a test of working memory (Sternberg, 1975). Subjects saw a set of five digits in a row on the screen. They were asked to memorize these five digits. Following that, a series of single digits appeared on the screen one by one. The task was to press the YES button if the digit was present in the memorized set or the NO button if it was not. Mean reaction time and percentage of errors were recorded.

2.1.5. Shape pairs (SP)

In this test of paired associate learning, subjects saw two shapes on the screen side by side (Tiplady et al., 2005). They were shown for three seconds. One of the shapes was then shown, and the task was to press the button on the side where that shape was seen. After a block of trials, subjects were shown a second pair of shapes, and then again responded to each shape depending on which side they had been shown; however, now any of the four shapes could appear, not just the first two. This was repeated until subjects had eight shapes to remember. Mean reaction time and percentage of errors were recorded.

2.1.6. Serial sevens (SS)

The serial sevens was a simple arithmetic test (Hayman, 1942). Subjects first saw a starter number in the range of 800–899, and then saw a series of 3-digit numbers decreasing in magnitude. The task was to tap 'yes' if the number was seven less than the previous number shown, and 'no' in all other cases. Mean reaction time and percentage of errors were recorded.

2.2. Statistical analysis

Data were analyzed using the Statistical Program for the Social Sciences (SPSS, Windows, IBM Corp), version 27. To analyze practice effects, scores from each training session and the final session were compared using GLM ANOVA for repeated measures for RTs, and the Related-Samples Friedman's two-way analysis of variance by ranks for % errors was conducted. If significant, F-tests were conducted to compare RTs of sessions 1–5 with session 6, and Related Samples Wilcoxon Signed Ranks tests were applied to compare percentage errors between sessions 1–5 and session 6, and Bonferroni's corrections were applied. To compare across platforms, for each training session and the final session, differences between the tablet and the phone were analyzed using a paired-samples *t*-test. All tests were two-tailed, and differences were regarded as significant if $p < 0.05$. Correlations between performance outcomes of the devices were considered significant if $p < 0.05$.

3. Results and discussion

3.1. Practice assessments

Data for the five practice assessments are shown in **Figure 2** and **Table 1**. In **Figure 2**, data are shown as differences from the value for assessment 6 (the index assessment). Significance was calculated for the difference from the assessment 6 value. Significant differences from the final assessment are indicated by the fill color of the markers: black: not significant; gray: $p < 0.05$; white: $p < 0.01$. For the reaction time data, there is a clear reduction in reaction time over the practice period for the majority of tests (see **Table 1**).

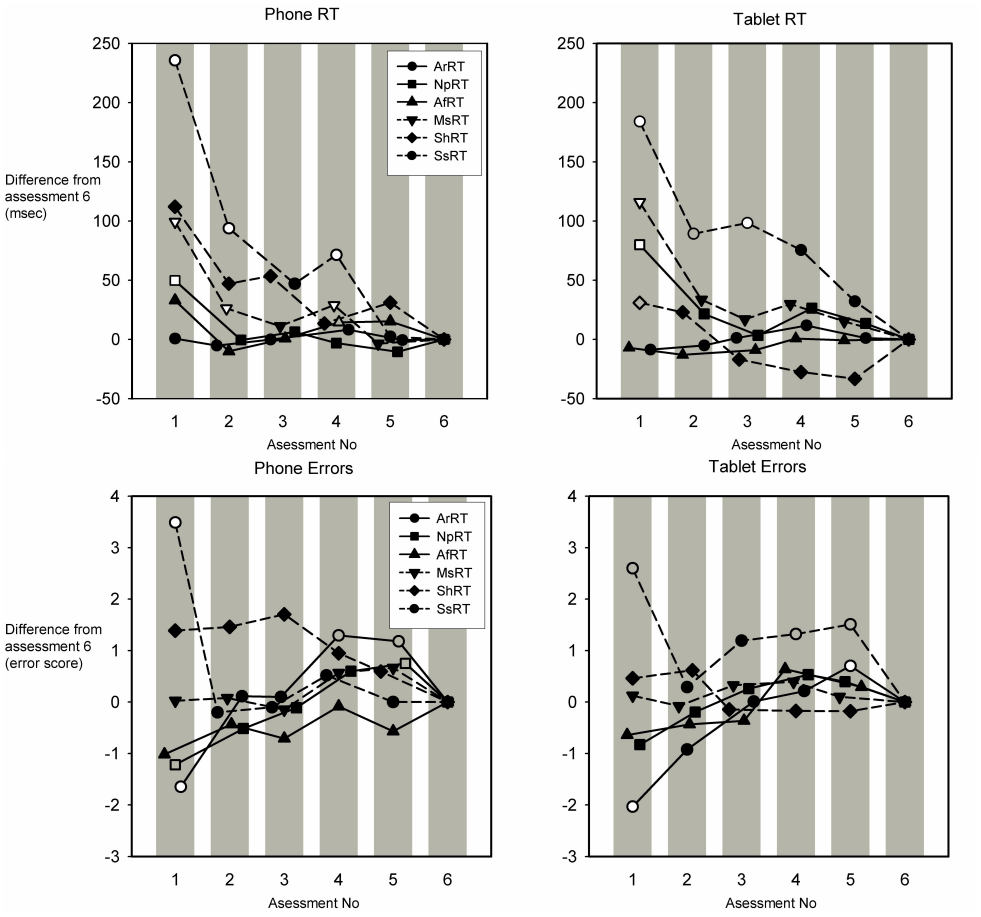

Figure 2. Practice effects for test scores on phone and tablet platforms.

Table 1. Reaction Time (RT) data.

Phones	**Overall**	**Session 1**	**Session 2**	**Session 3**	**Session 4**	**Session 5**	**Session 6**
AR	416.0 (34.2)	415.8 (30.0)	410.1 (32.7)	415.6 (32.4)	423.9 (38.1)	415.5 (56.7)	417.0 (47.5)
***p*-value**	0.188	-	-	-	-	-	
NP	659.7 (80.0)	721.0 (86.2)	670.4 (66.6)	677.7 (72.4)	668.8 (88.4)	661.2 (74.1)	661.7 (100.9)
***p*-value**	<0.0001	<0.0001	n.s.	n.s.	n.s.	n.s.	
AF	541.3 (63.3)	577.6 (117.0)	538.1 (69.5)	547.8 (67.3)	560.8 (80.5)	561.4 (102.6)	542.1 (83.7)
***p*-value**	0.048	n.s.	n.s.	n.s.	0.021	n.s.	
MS	687.5 (93.1)	796.4 (139.5)	727.5 (101.5)	709.3 (113.9)	726.4 (113.7)	694.8 (100.8)	688.6 (101.8)
***p*-value**	<0.0001	<0.0001	0.001	n.s.	<0.0001	n.s.	
SP	740.2 (172.8)	833.2 (178.4)	771.2 (140.2)	772.5 (180.5)	733.7 (140.0)	751.4 (160.1)	738.2 (151.2)
***p*-value**	0.061	-	-	-	-	-	
SS	776.1 (197.3)	1022.8 (350.4)	885.7 (238.7)	837.5 (176.7)	863.2 (217.5)	794.7 (205.0)	773.1 (216.4)
***p*-value**	<0.0001	<0.0001	<0.0001	0.032	0.002	n.s.	
Tablets	**Overall**	**Session 1**	**Session 2**	**Session 3**	**Session 4**	**Session 5**	**Session 6**
AR	414.0 (59.0)	406.7 (35.9)	409.9 (38.7)	416.3 (48.5)	427.1 (65.5)	415.8 (54.5)	414.1 (58.2)
***p*-value**	0.265	-	-	-	-	-	
NP	679.6 (119.8)	753.5 (113.2)	693.6 (106.8)	677.2 (94.8)	699.4 (147.2)	684.3 (120.2)	681.1 (118.5)
***p*-value**	0.001	<0.0001	n.s.	n.s.	n.s.	n.s.	
AF	551.1 (102.1)	537.6 (51.1)	533.0 (72.6)	535.8 (67.2)	545.6 (93.8)	544.1 (99.6)	549.6 (101.2)
***p*-value**	0.340	-	-	-	-	-	

Table 1. *Cont.*

Tablets	Overall	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6
MS	706.9 (112.4)	815.5 (138.3)	731.9 (110.8)	714.5 (108.1)	725.2 (129.5)	714.2 (111.8)	707.6 (111.0)
***p*-value**	<0.0001	<0.0001	n.s.	n.s.	n.s.	n.s.	
SP	702.9 (129.8)	757.2 (131.4)	741.7 (119.7)	702.9 (146.2)	692.6 (136.7)	689.3 (125.7)	703.6 (128.2)
***p*-value**	0.034	0.037	n.s.	n.s.	n.s.	n.s.	
SS	811.8 (237.0)	974.4 (315.4)	881.0 (203.8)	885.9 (226.9)	863.1 (275.2)	821.1 (204.6)	805.7 (236.9)
***p*-value**	<0.0001	<0.0001	0.029	0.002	n.s.	n.s.	

Notes: Mean (SD) RT are shown. Related-Samples Friedman's two-way analysis of variance by ranks tests were conducted to examine whether there was an overall significant difference between the sessions. If significant ($p < 0.05$), Related Samples Wilcoxon Signed Ranks tests were conducted to compare the individual sessions 1–5 with session 6. Abbreviations: n.s. = not significant, - = not conducted, AF = Arrow Flankers, AR = Arrow Reaction time, NP = Number Pairs, MS = Memory Scanning, SP = Shape Pairs, SS = Serial Sevens.

GLM ANOVA for repeated measures revealed no overall difference between the test sessions for AR and SP on phones, and AR and AF on tablets. For these test outcomes, there was no suggestion of improvement after the first practice assessment. For tests other than AR, significant improvement was seen over the practice period. All other tests showed significantly higher values for RT at assessment 1 compared to the final assessment, with varying rates of improvement thereafter. No test showed a significant difference between assessment 5 (the last practice trial) and the final assessment.

For the error data (see **Table 2**), the pattern is much less clear-cut, with both increases and decreases in performance seen. Related-Samples Friedman's two-way analysis of variance by ranks tests revealed no overall difference between the test sessions for AF, MS, and SP on phones, and NP, AF, MS, and SP on tablets. Significant practice effects were seen for AR, NP, and SS. For both phone and tablet, Serial Sevens (SS) showed a significantly higher error rate at the first assessment, while AR showed a significantly lower rate at this timepoint. Several individual significant comparisons were found after this, but no consistent pattern of change was seen for time-points 2 to 5 (see **Table 2**).

Table 2. Percentage errors (%).

Phones	Overall	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6
AR	4.2 (3.3)	3.2 (2.9)	4.7 (3.3)	5.1 (3.4)	5.4 (3.9)	5.6 (3.7)	4.6 (3.5)
***p*-value**	<0.0001	0.001	n.s.	n.s.	0.044	0.025	
NP	4.3 (3.2)	3.7 (2.7)	4.4 (2.7)	4.6 (2.8)	5.4 (3.5)	5.1 (3.1)	4.6 (3.2)
***p*-value**	<0.0001	0.021	n.s.	n.s.	n.s.	0.013	
AF	2.1 (2.5)	2.4 (1.2)	3.1 (1.6)	2.5 (1.4)	3.2 (2.3)	3.1 (1.9)	3.0 (2.1)
***p*-value**	0.137	-	-	-	-	-	
MS	3.5 (2.3)	3.7 (2.0)	4.2 (1.9)	3.6 (2.5)	4.4 (2.3)	4.5 (3.2)	3.8 (2.2)
***p*-value**	0.458	-	-	-	-	-	
SP	2.8 (2.6)	6.2 (5.4)	6.8 (5.1)	6.7 (4.7)	5.6 (3.0)	4.9 (3.3)	4.0 (2.7)
***p*-value**	0.640	-	-	-	-	-	

Table 2. Percentage errors (%).

Phones	**Overall**	**Session 1**	**Session 2**	**Session 3**	**Session 4**	**Session 5**	**Session 6**
SS	3.3 (4.6)	7.8 (5.9)	4.7 (2.2)	5.2 (4.3)	4.9 (5.7)	6.0 (5.7)	5.0 (4.7)
***p*-value**	0.002	0.002	n.s.	n.s.	n.s.	n.s.	
Tablets	**Overall**	**Session 1**	**Session 2**	**Session 3**	**Session 4**	**Session 5**	**Session 6**
AR	4.0 (3.8)	2.9 (1.8)	3.4 (2.3)	4.8 (3.2)	4.9 (3.1)	5.2 (3.5)	4.5 (3.8)
***p*-value**	<0.0001	<0.0001	n.s.	n.s.	n.s.	0.004	
NP	4.5 (3.4)	3.9 (2.5)	4.6 (2.7)	5.0 (2.7)	5.4 (3.0)	5.0 (3.3)	4.8 (3.4)
***p*-value**	0.115	-	-	-	-	-	
AF	2.6 (1.9)	3.3 (2.8)	2.7 (1.7)	3.1 (2.0)	3.7 (2.3)	3.6 (2.4)	3.1 (1.7)
***p*-value**	0.005	-	-	-	-	-	
MS	3.0 (2.4)	4.4 (2.3)	3.9 (2.0)	4.1 (2.1)	4.9 (2.2)	4.3 (1.9)	3.8 (2.2)
***p*-value**	0.372	-	-	-	-	-	
SP	3.6 (3.5)	5.8 (4.2)	5.0 (4.9)	5.0 (2.9)	4.2 (2.5)	4.5 (3.9)	5.2 (3.0)
***p*-value**	0.962	-	-	-	-	-	
SS	3.5 (4.5)	7.5 (4.9)	5.4 (3.7)	6.1 (5.3)	6.3 (3.9)	6.0 (4.9)	5.3 (4.7)
***p*-value**	0.027	0.017	n.s.	n.s.	0.044	0.032	

Notes: Mean (SD) % errors are shown. Related-Samples Friedman's two-way analysis of variance by ranks tests were conducted to examine whether there was an overall significant difference between the sessions. If significant ($p < 0.05$), Related Samples Wilcoxon Signed Ranks tests were conducted to compare the individual sessions 1-5 with session 6. Abbreviations: n.s. = not significant, - = not conducted, AF = Arrow Flankers, AR = Arrow Reaction time, NP = Number Pairs, MS = Memory Scanning, SP = Shape Pairs, SS = Serial Sevens.

Given that the reaction time practice data for four tests were consistent in showing significant improvement over the practice period, the mean RT scores and significance comparisons for these tests were calculated for phone and tablet. Data are shown in **Figure 3**. Patterns are very similar for the two platforms, and suggest that 3 or 4 practice assessments for these tests are sufficient to ensure stable performance.

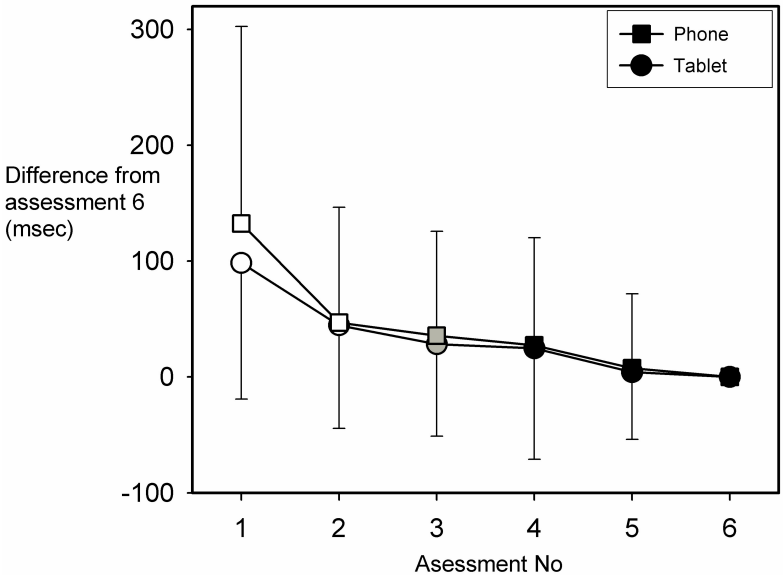

Figure 3. Mean practise effects.

Notes: Mean practice effect scores for the four tests (NP, MS, SP, and SS). Significant differences from the final assessment are indicated by fill color: black: not significant; gray: $p < 0.05$; white: $p < 0.01$.

3.2. Comparison between devices, final assessment

Scores for test performance on the two platforms have been compared in two ways. Firstly, the mean scores for the final assessment have been compared, and the difference expressed as an effect size (mean difference/pooled standard deviation). Secondly, the correlation between scores for phone and tablet administration has been calculated. Results are shown in **Table 3**.

Table 3. Comparison of test scores for phone and tablet administration.

Variable	Mean (SD)	*t*-value	Probability	Effect size	Correlation phone vs tablet
AR RT	2.0 (48.2)	0.244	0.809	0.041	0.536
AR % Errors	0.2 (3.5)	0.574	0.569	0.067	0.756
NP RT	−19.8 (101.9)	−1.526	0.135	−0.195	0.748
NP % Errors	−0.3 (3.3)	−0.638	0.527	−0.077	0.726
AF RT	−9.9 (85.0)	−0.994	0.327	−0.117	0.824
AF % Errors	−0.5 (2.2)	−1.459	0.153	−0.233	0.530
MS RT	−19.4 (103.2)	−1.833	0.075	−0.189	0.814
MS % Errors	0.5 (2.3)	1.622	0.113	0.224	0.640
SP RT	37.3 (152.9)	1.642	0.109	0.244	0.605
SP % Errors	−0.7 (3.1)	−1.134	0.264	−0.241	0.144
SS RT	−35.6 (218.0)	−1.309	0.198	−0.164	0.716
SS % Errors	−0.1 (4.6)	−0.236	0.815	−0.028	0.751

Notes: Mean (SD) difference scores are presented. Difference scores represent the difference between the overall mean scores (sessions 1–6) of phones minus tablets. Abbreviations: RT = reaction time, AF = Arrow Flankers, AR = Arrow Reaction time, NP = Number Pairs, MS = Memory Scanning, SP = Shape Pairs, SS = Serial Sevens.

In no case was the difference between test scores between phone and tablet statistically significant. Effect sizes range from 0.028 to 0.244. Cohen has suggested that an effect size of 0.2 between two means should be considered small, 0.5 medium, and 0.8 large (Cohen, 1992). Thus, all the effect sizes seen here would be considered small. There is no clear trend for performance to be better on one platform than the other. With one exception, correlations between phone and tablet scores were >0.5, and in the majority of cases, >0.7, indicating reasonable to good agreement between platforms. All correlations were statistically significant ($p < 0.05$), with the exception of the error score for SP.

3.3. Mental effort scale

No significant differences in mental effort were observed between the subsequent sessions or between the devices (see **Table 4**).

Table 4. Mean (SD) perceived mental effort after the session per device.

Device	Training 1	Training 2	Training 3	Training 4	Training 5	Final session
Phone	5.2 (2.1)	5.7 (2.2)	5.3 (2.0)	5.4 (2.4)	5.1 (2.5)	5.7 (2.6)
Tablet	5.2 (1.9)	5.6 (2.3)	5.1 (2.3)	5.5 (2.4)	4.8 (2.5)	5.6 (2.7)

4. Discussion

Since the 1990s, portable performance platforms based on mobile technology have been increasingly used in clinical settings for cognitive and psychomotor testing (Tiplady, 1994). Not only do portable platforms make it easier to assess many subjects at one time, but using these devices also realizes the possibility to perform cognitive testing outside the laboratory, or online testing at home. The latter is important if one wishes to test subjects in their natural environment, or test many subjects at the same time using their own handheld device without interference from a researcher (Waters et al., 2012).

The current study examined the usefulness of six mobile tests to be used in future psychopharmacological research, comparing them between platforms with very different screen sizes. Another important difference between the devices was the fact that the phones had a keypad and the tablets had a touchscreen. Despite these differences, the results indicated that the scores were similar when obtained from tests administered on small-screen phone platforms or larger-screen tablet platforms, or between keypad and touchscreen. Mean scores for the final index assessment were similar on both platforms, with no significant differences between platforms and small effect sizes for the observed mean differences. Patterns of scores over the practice period were also similar between the two platforms. Most modern smartphones have a larger screen size than the phones used in the current study. It is therefore reasonable to assume that performance on phones with a larger screen size, i.e., intermediate between the two devices used here, will also not differ from performance on tablets.

Comparability between platforms was also addressed by evaluating the correlations between scores for phone and tablet test scores. Most correlations were greater than 0.5, indicating reasonable agreement, and the majority were over 0.7, indicating good agreement (Cohen, 1992). The one exception was the error score for the Shape Pairs task. There is no obvious reason for this discrepancy, as there was no indication of instability of this measure from other indices. Further work is needed to address this.

A limitation of this analysis is that the study design did not permit issues of between-mode and within-mode agreement to be addressed. Thus, a low correlation could be due to the test measure itself being unreliable, or to an issue between the two platforms. For the measures showing satisfactory correlations, this is less of a problem, as the within-platform reliability must be at least as great as the observed between-platform agreement.

An important practical issue is the number of practice/training assessments needed to establish stable levels of performance. For the test battery as a whole, it is recommended to have a minimum of three practice administrations to ensure stable performance, irrespective of whether phone or tablet platforms are used. In the current study design, we allowed extensive time to recover from the test battery to rule out possible fatigue or motivational effects. In research practice, this time could be significantly shortened, or if a study design allows, the training could be divided over multiple screening days. Important in this context is that the comparison between

the devices was conducted in a sample of young and healthy volunteers. Also of note, the results are presented at the group level. Therefore, in practice, individual participants may require more practice sessions to attain baseline performance levels. For example, participants, such as the elderly, patients with cognitive disorders or learning disabilities, or people undergoing psychopharmacological treatment, may require additional training sessions to attain baseline performance on the tests.

No significant differences were observed in perceived effort to complete the tests on a phone or tablet. Differences in perceived effort were also not observed between the test sessions, indicating that the scheduled breaks in the study design were of sufficient duration to recover from the previous test session. In the current study, we did not compare different types of tablets or phones. However, when testing in naturalistic settings, participants may wish to download the tests and complete them on their own device. Devices may differ in layout, keyboard, and touch sensitivity. Therefore, in future studies, the different types and popular brands of devices should also be compared. As we examined the extremes in terms of screen size in the current study, we are confident that the mobile test battery can be successfully implemented on such devices of intermediate screen size.

5. Conclusion

Taken together, this study confirmed that the test battery under investigation can be administered both on phones and tablets, and that comparable results will be obtained. However, training sessions are recommended to achieve stable performance levels.

Author contributions: Conceptualization, JCV, JJ, AS, BT, JG, and AJvdL; methodology, BT; software, BT; formal analysis, BT; investigation, JJ; writing—original draft preparation, JCV; writing—review and editing, JCV, JJ, AS, BT, JG, and AJvdL. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional review board statement: The study was conducted in accordance with the guidelines of the Declaration of Helsinki and its latest amendments. The University of Groningen Psychology Ethics Committee approved the study (approval code: ppo-014-224, approval date: 30 April 2015).

Informed consent statement: Written informed consent was obtained from all participants involved in the study.

Data availability statement: The data are available upon reasonable request from the corresponding author.

Conflict of interest: Johan Garssen is a part-time employee of Nutricia Research and received research grants from Nutricia Research Foundation, Top Institute Pharma, Top Institute Food and Nutrition, GSK, STW, NWO, Friesland Campina, CCC, Raak-Pro, and the EU. Over the past 36 months, J.V. has acted as a consultant/expert advisor to Eisai, KNMP, Med Solutions, Mozand, Red Bull, Sen-Jam Pharmaceutical, and Toast! J.V. has received travel support from Sen-Jam Pharmaceutical and owns stock

in Sen-Jam Pharmaceutical. A.S. has acted as a consultant/expert advisor to Bayer, Coca Cola, Danone, Delica Therapeutics, GlaxoSmithKline, Givaudin, Liquid IV, Mars-Wrigley, Naturex, Nestlé, McCormick, Metavate Consultancy, PepsiCo, Pfizer, Pharmavite, REVIV, Sanofi, Verdure Sciences, and Wörwag Pharma; and in the past 36 months Scholey has held research grants from Abbott Nutrition, Arla Foods, the Australian Research Council, Bayer, BioRevive, DuPont, Fonterra, GlaxoSmithKline, the High Value Nutrition Fund, the National Health and Medical Research Council, Nutricia-Danone, Sanofi and Wörwag Pharma. A.S. is on the Scientific Advisory Board of Sen-Jam Pharmaceutical. He has stock from Sen-Jam Pharmaceutical and Ārepa Nootropics. He has received travel support from Vitafoods and ILSI. Brian Tiplady is the owner of Mobile Cognition Ltd, which developed the Penscreen test battery, and holds shares in AstraZeneca. The other authors have no conflicts of interest.

References

Bauer, R. M., Iverson, G. L., Cernich, A. N., et al., 2012. Computerized neuropsychological assessment devices: joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. Clinical Neuropsychology, 26, 177–196. DOI: https://doi.org/10.1093/arclin/acs027

Cohen, J. A., 1992. A power primer. Psychological Bulletin, 112, 155–159. DOI: https://doi.org/10.1037//0033-2909.112.1.155

Collie, A., Maruff, P., Darby, D. G., et al., 2003. The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. Journal of the International Neuropsychology Society, 9, 419–428. DOI: https://doi.org/10.1017/S1355617703930074

Eriksen, B. A., Eriksen, C. W., 1974. Effects of noise letters upon the identification of a target letter in a nonsearch task. Perception and Psychophysics, 16, 143–149. DOI: https://doi.org/10.3758/BF03203267

Hayman, M., 1942. Two minute clinical test for measurement of intellectual impairment in psychiatric disorders. Archives of Neurology and Psychiatry, 47, 454–464. DOI: https://doi.org/10.1001/archneurpsyc.1942.02290030112010

Jones, A., Tiplady, B., Houben, K., et al., 2018. Do daily fluctuations in inhibitory control predict alcohol consumption? An ecological momentary assessment study. Psychopharmacology, 235, 1487–1496. DOI: https://doi.org/10.1007/s00213-018-4860-5

Keenan, E. K., Tiplady, B., Priestley, C. M., et al., 2014. Naturalistic effects of five days of bedtime caffeine use on sleep, next-day cognitive performance, and mood. Journal of Caffeine Research, 4, 13–20. DOI: https://doi.org/10.1089/jcr.2011.0030

Lamond, N., Dawnson, D., Roach, G., 2005. Fatigue assessment in the field: validation of a hand-held electronic psychomotor vigilance task. Aviation, Space, and Environmental Medicine, 76(5), 486–489.

Meyers, C. A., Brown, P. D. J., 2006. Role and relevance of neurocognitive assessment in clinical trials of patients with CNS tumors. Journal of Clinical Oncology, 24, 1305–1309. DOI: https://doi.org/10.1200/JCO.2005.04.6086

Mulder-Hajonides van der Meulen, W., Wijnberg, J., Hollander, J., et al., 1980. Measurement of subjective sleep quality. In: Proceedings of the International European Sleep Congress; 2–5 September 1980; Amsterdam, The Netherlands.

Sliwinski, M. J., Mogle, J. A., Hyun, J., et al., 2018. Reliability and validity of ambulatory cognitive assessments. Assessment, 25, 14–30. DOI: https://doi.org/10.1177/1073191116643164

Sternberg, S., 1975. Memory scanning: New findings and current controversies. Quarterly Journal of Experimental Psychology, 27, 1–32. DOI: https://doi.org/10.1080/14640747508400459

Sternin, A., Burns, A., Owen, A. M., 2019. Thirty-five years of computerized cognitive assessment of aging—Where are we now? Diagnostics, 9, 114. DOI: https://doi.org/10.3390/diagnostics9030114

Stone, A. A., Shiffman, S., 1994. Ecological momentary assessment (EMA) in behavorial medicine. Annals of Behavioral Medicine, 16, 199–202. DOI: https://doi.org/10.1093/abm/16.3.199

Thomson, A. J., Nimmo, A. F., Tiplady, B., et al., 2009. Evaluation of a new method of assessing depth of sedation using two-choice visual reaction time testing on a mobile phone. Anaesthesia, 64, 32–38. DOI: https://doi.org/10.1111/j.1365-2044.2008.05683.x

Tiplady, B., 1994. The use of personal digital assistants in performance testing in psychopharmacology. In: Proceedings of the British Psychological Society Annual Conference; 24–27 March 1994; Brighton, UK.

Tiplady, B., Degia, A., Dixon, P., 2005. Assessment of driver impairment: Evaluation of a two-choice tester using ethanol. Transportation Research Part F: Traffic Psychology and Behavior, 8, 299–310. DOI: https://doi.org/10.1016/j.trf.2005.04.013

Verster, J. C., Tiplady, B., McKinney, A., 2012. Mobile technology and naturalistic study designs in addiction research. Current Drug Abuse Reviews, 5(3), 169–171. DOI: https://doi.org/10.2174/1874473711205030169

Vincent, A. S., Bailey, C. M., Cowan, C., et al., 2017. Normative data for evaluating mild traumatic brain injury with a handheld neurocognitive assessment tool. Applied Neuropsychology: Adult, 24(6), 566–576. DOI: https://doi.org/10.1080/23279095.2016.1213263

Waters, A. J., Marhe, R., Franken, I. H. J. P., 2012. Attentional bias to drug cues is elevated before and during temptations to use heroin and cocaine. Psychopharmacology, 219, 909–921. DOI: https://doi.org/10.1007/s00213-011-2424-z

Zijlstra, F. R. H., van Doorn, L., 1985. The construction of a scale to measure perceived effort. Delft: Delft University of Technology.