Psychological Test and Assessment Modeling, Volume 63, 2021 (4), 507-524

Are you Swiping, or Just Marking? Exploring the Feasibility of Psychological Testing on Mobile Devices

Marco Koch[1,2], Corina Möller[3], Frank M. Spinath[2]

Abstract

Despite the many benefits of computer-based testing, many existing computer-based tests employ response formats that could be used equally in paper-pencil tests. In this article we explore the feasibility of psychological testing on mobile devices, an approach that combines the advantages of computer-based testing with the flexibility of paper-pencil tests. As an example, we present the Attention Swiping Task (AST) for assessing sustained attention on mobile devices in proctored or self-administered settings. N = 114 university students were tested with the AST, another test measuring sustained attention (FAIR-2), and a figural matrices test (DESIGMA) measuring participants' intelligence (IQ) to evaluate the psychometric properties and construct validity of the AST. Results indicated that the AST had a satisfying distribution of item difficulties (M_{Diff} = .58, SD_{Diff} = .36) and part-whole correlations (M_{PWC} = .55, SD_{PWC} = .11), and an excellent reliability (r_{tt} = .99). Moreover, test indices of the AST were highly correlated with the FAIR-2. Furthermore, there were small significant positive correlations between AST test indices and participants' IQ (r = .23 - .25, p = .04 - .02). These results indicate that the AST can be reliably applied for measuring sustained attention by means of mobile devices. Moreover, in contrast to existing tests of sustained attention the AST can be customized easily, is applicable for (self-administered) online studies, and its source code is released freely under the GNU GPLv3 license. This also serves as foundation for the development of further psychological tests for mobile devices.

Keywords: sustained attention, computer-based test (CBT), test development and evaluation, cognitive abilities, mobile device adoption

[1] Correspondence concerning this article should be addressed to: Marco Koch, Campus A1.3, 66123 Saarbrücken, Germany, marco.koch@uni-saarland.de

[2] Individual Differences & Psychodiagnostics, Saarland University, Germany

[3] Developmental Psychology Unit, Saarland University, Germany

1. Introduction

In the past decade, the advantages of computer-based tests (CBTs) have been increasingly acknowledged (see Tippins, 2015 for a review), and a large number of CBTs was implemented in various study designs (Ghaderi et al., 2015). At the beginning of 2020, due to the COVID-19 pandemic, researchers were suddenly required to carry out most of their studies online. This situation further emphasized the need for CBTs in various psychological research communities.

To date, there is evidence that attitudes and personality traits can be reliably assessed by both CBTs and paper-pencil tests (Tippins, 2015), whereas evidence regarding ability tests is less consistent. An early meta-analysis (Mead & Drasgow, 1993) found that differences between CBTs and paper-pencil tests mainly occurred when ability tests were speeded (see also Potosky & Bobko, 2004). More recent studies on educational tests (e.g., tests of mathematical proficiency) supported the notion of measurement invariance (Kingston, 2008; Wang et al., 2007, 2008), but also identified several factors that can cause differences between CBTs and PBTs (e.g., familiarity with using computers, conventional tests vs. adaptive tests).

It is noteworthy that most CBTs were developed as an adaptation of existing paper-pencil tests, facilitating test construction and validation procedures considerably. At the same time this approach prevents test developers from utilizing the full potential of CBTs (e.g., customization of instructions, stimuli or time limits, comparisons of test behavior at the beginning, middle, and end of a test session). In the context of intelligence assessment, this approach has been called *computerized pen-and-paper tests*, and it has been proposed that future research and test development should focus on innovative test formats that are explicitly developed for computerized administration (Koch, Becker, et al., 2021). To date, there is little research regarding innovative approaches to testing and their feasibility. One special form of computer-based testing is the utilization of mobile devices. In the last decade, the usage of mobile devices such as smartphones or tablet computers has dramatically increased in developed countries, and in developing countries people have better access to smartphones than computers (Joshi & Avasthi, 2007). Mobile devices inherently offer advantages over traditional computers because they are more affordable and provide very intuitive modes of interaction (i.e., touching user interface elements with one’s finger instead of using a computer mouse). To date, empirical studies that compared psychometric tests administered on traditional computers versus mobile devices have reported mixed results (Arthur et al., 2014; Brown & Grossenbacher, 2017; King et al., 2015). However, in all these studies, multiple-choice procedures which easily fall into the category of *computerized pen-and-paper tests* were implemented and investigated. In contrast, the present study explores the feasibility of an innovative test format for mobile devices which was developed to measure sustained attention, using a novel response format that is native to mobile devices.

Sustained attention (sometimes also referred to as concentration; Blotenberg & Schmidt-Atzert, 2019a) is the ability to direct one’s attentional focus to specific stimuli in a relevant task and to maintain this state over a prolonged period of time (Schweizer, 2005; Williams & Saunders, 1997). This ability is essential for mastering various academic, professional and everyday life activities, and is regarded as a prerequisite of higher cognitive processes (e.g., working memory or reasoning; Burack et al., 2012; Krumm et al., 2012; Lezak, 1995; Schumann, 2015). Therefore, tests assessing sustained attention are frequently applied in a broad range of psychological and educational disciplines (Blotenberg & Schmidt-Atzert, 2019b; Moosbrugger & Goldhammer, 2006; Schmidt-Atzert et al., 2006).

Most tests of sustained attention follow Bourdon's (1895) basic construction principles in such ways that they consist of rows containing relevant (target) and irrelevant (non-target) stimuli. Within a given time limit, test-takers are required to respond to (e.g., by ticking off, crossing out or clicking on an item) as many targets as possible while refraining from responding to non-targets. Typically, tests of sustained attention employ simple and homogenous stimuli (e.g., letters, numbers, or simple figures), and implement rather simple mental operations (Büttner & Schmidt-Atzert, 2004; Westhoff & Hagemeister, 2005). To date, several CBTs of sustained attention exist (e.g., d2-R, FACT-2, and SART: Brickenkamp et al., 2010; Goldhammer et al., 2009; Robertson et al., 1997), and there is evidence for the comparability with their respective paper-pencil versions (Krumm et al., 2008, 2012). Nonetheless, tests of sustained attention are still mainly administered as paper-pencil tests

To fill this gap, we designed a CBT of sustained attention for the application on mobile devices. The Attention Swiping Task (AST) implements traditional construction principles (rows of target and non-target stimuli; simple mental task) as proposed by Bourdon (1895), a complete cancellation procedure (e.g., Moosbrugger et al., 2011), and a self-paced response mode with a fixed time limit of three minutes. The stimulus material consists of pictorial flowers with either two or three red petals, and with either a blue square or circle in the center. Each combination appears in one of four orientations resulting in 16 distinct stimuli (***Figure 1 A***). In total, the test encompasses 720 items with an equal number of targets and non-targets. Stimuli were randomly grouped into rows of nine items resulting in 80 test rows. In the center of each test page, a row of nine items surrounded by a black frame is presented (Figure 1 B). It is noteworthy that all test features (stimuli, time limit, ratio of targets and distractors) can be customized (e.g., task switching version of the AST, Hofer et al., in press).

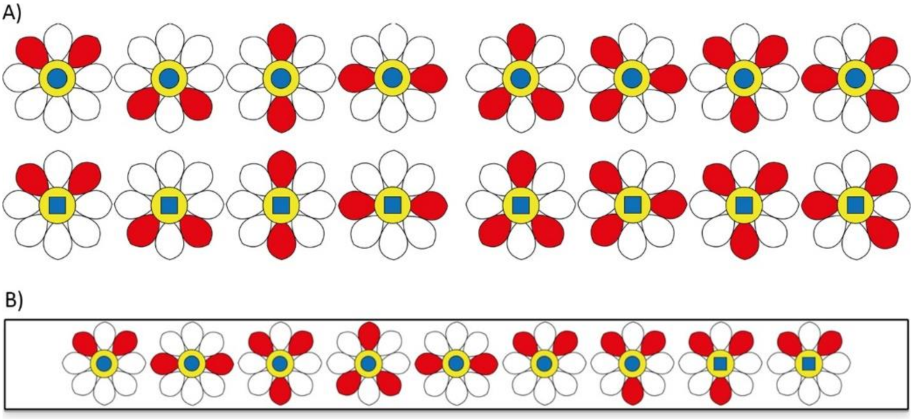

Figure 1. A) All 16 Stimuli implemented in the AST. B) Screenshot of a test row as presented in the AST.

First, test takers are instructed to memorize two rules (e.g., 1. items with two red petals and a blue circle as well as items with three red petals and a blue square need to be swiped up; 2. all other items need to be swiped down). Subsequently, test takers are familiarized with the task by completing a practice row. Test takers are instructed to always respond from left to right, with the next item being locked until the previous item is responded to. During the familiarization period, feedback is given (indicated by red and green frames appearing around the item for incorrect and correct responses, respectively). Incorrect responses prompt a pop-up window providing an explanation why the response was incorrect and requesting the test taker to correct the response. If test takers commit six or more errors in a practice row, another practice row appears. This procedure is repeated up to three times. If test takers respond incorrectly to six or more items in the last practice row, they can nonetheless continue to the actual test. However, the number of practice rows and errors are logged and can be used for exclusion. After passing the familiarization period, the actual test begins, and test takers are instructed to respond as quickly as possible with a time limit of three minutes. Once a row is completed, a new row of nine items appears in the middle of the screen. During the test phase, no feedback is given, and responses cannot be corrected.

The aim of the present research was to evaluate the psychometric properties of the AST regarding its reliability, construct validity and criterion validity. Furthermore, it was investigated whether the AST showed a one-factorial structure, comparable to other tests of sustained attention. To this end, the AST was investigated in two studies. In the first study, the AST was administered to a student sample in a proctored lab setting, whereas in study 2, a heterogeneous adult sample was investigated in a field setting without experimenters' guidance.

2. Study 1

Study 1 aimed to explore the feasibility of psychological tests for mobile devices by administering the AST and evaluating its psychometric properties in a student sample. Moreover, it was investigated whether individual differences in the attitude towards technology and the generation of the respective mobile device had an impact on the test scores of the AST.

2.1 Method

2.1.1 Sample

A total of 117 psychology students participated. After excluding three participants for not following the instructions of the AST, the final sample consisted of *n* = 114 participants (*n* = 79 female, *n* = 30 male, *n* = 5 did not provide information about their gender) with a mean age of 24.97 years (*SD* = 8.91). In the final sample, 10 % of the participants (*n* = 11) were left-handed and *n* = 1 participant indicated to have dyschromatopsia. Participants received course credit for their participation.

2.1.2 Measurement instruments

To assess convergent validity, the paper-pencil version of the FAIR-2 (Moosbrugger & Oehlschlägel, 2011) was administered. The FAIR-2 has a high internal consistency (r_{tt} > .90) and satisfying retest reliability within an interval of two weeks (r_{tt} = .81). Convergent validity for the FAIR-2 was demonstrated by a mean correlation of *r* = .49 with other tests of sustained attention (e.g., d2; Brickenkamp, 1981). Administration of the FAIR-2 takes 10-12 minutes.

To assess divergent validity, a short form of the DESIGMA (Becker & Spinath, 2014) was used. Participants were required to solve 12 figural matrices with increasing difficulty. DESIGMA has a high internal consistency (α > .91).

In addition, affinity towards technology was assessed by means of the TA-EG (Karrer et al., 2009). This questionnaire consists of 19 items measuring four scales (enthusiasm, competency, positive attitudes, negative attitudes) with satisfying reliability coefficients (.73 < α < .86).

2.1.3 Procedure

First, the paper-pencil version of the FAIR-2 (Moosbrugger & Oehlschlägel, 2011) was administered to all participants in a group setting. Upon completion, all further measurements were administered on the participants’ personal mobile devices. Participants were required to enter sociodemographic information, the generation of their mobile device, and whether they suffered from dyschromatopsia. Subsequently, the AST was administered in a self-paced mode, followed by the DESIGMA (Becker & Spinath, 2014). Finally, participants were required to complete the TA-EG (Karrer et al., 2009).

2.1.4 Statistical Analyses

Analyses were carried out in *R* statistics (R Core Team, 2021). Item difficulties were calculated as one minus the relative frequency of an item being solved. Part-whole correlations were computed with the R package *psych* (Revelle, 2017). However, since the AST is a speeded test – with a large number of items to be solved within three minutes – item difficulties and part-whole correlations were aggregated by calculating means of all nine items per test page.

Test reliability coefficient was calculated by the odd-even split-half approach. To test for the proposed one-factorial structure of the AST, a confirmatory factor analysis (CFA) was carried out with the R package *lavaan* (Rosseel, 2012). Because single items of a speeded test are redundant and prone to random errors, a parceling approach as described by Matsunaga (2008) was applied. To evaluate the model fit, we used the indicators CFI (CFI > .95 indicates good model fit; Hu & Bentler, 1999) and the SRMR (SRMR < .08 indicates good model fit; Hu & Bentler, 1999).

Prior to correlational analyses, performance indices as described for the FAIR-2 (Moosbrugger & Oehlschlägel, 2011) were calculated based on the raw response data of the AST. These comprise an index of attentive performance (L), an index of quality (Q), and an index of continuity (K). Construct validity was analyzed using bivariate correlations between AST and FAIR-2 (Moosbrugger & Oehlschlägel, 2011), and AST and DESIGMA (Becker & Spinath, 2014), respectively. Associations between the performance indices and attitudes towards technology were also analyzed using bivariate correlations with the four scales of the TA-EG.

To test the impact of the generation of the used hardware, a multivariate ANOVA with the factor *mobile device generation* (newer than one year, one to two years old, two to three years old, three to four years old, four to five years old, and older than five years) was conducted for the three performance indices L, Q, and K.

2.2 Results

On average, participants attempted 137.5 items (*SD* = 31.79) of which they solved 130.35 items (*SD* = 32.85) correctly. Results of the item parameter analyses are depicted in **Figure 2**. Analyses revealed a mean item difficulty of M_{Diff} = .58 (SD_{Diff} = .36) and a part-whole correlation coefficient of M_{PWC} = .55 (SD_{PWC} = .11). Test reliability coefficient was r_{tt} = .99. The CFA revealed an excellent model fit (GFI = 1.00, SRMR = .004), and the latent reliability was comparable to the manifest reliability (ω = .99).

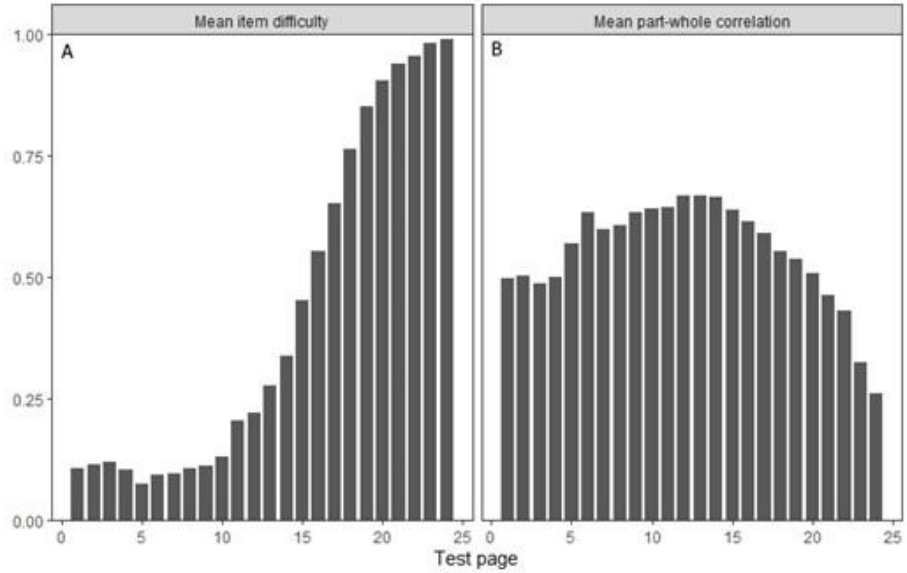

Figure 2.

A) Mean item difficulties per test page. B) Mean part-whole correlations per test page.

Bivariate correlation analyses revealed that the performance indices G, L, and K of the AST and FAIR-2 were significantly correlated (G index: r_G = .53, p_G < .001; L index: r_L = .47, p_L < .001; K index: r_K = .43, p_K < .001, see Table 1). No significant correlations were found between the error indices F_V and $F_{A,}$ and the Q index of the AST and the FAIR-2 (*all R^2 < .02*).

Table 1.

Correlations between test scores of AST and FAIR-2

	AST G	AST FV	AST FA	AST L	AST Q	AST K
FAIR-2 G	.53*	.05	-.07	.47*	.09	.40*
FAIR-2 FV	-.04	.05	.14	-.10	-.12	.13
FAIR-2 FA	.04	.14	.14	.05	-.16	-.09
FAIR-2 L	.51*	.01	-.12	.47*	.13	.42*
FAIR-2 Q	.14	-.07	-.15	.19*	.14	.21*
FAIR-2 K	.49*	-.02	-.15	.47*	.16	.43*

Notes. G = number of items attempted; FV = misses; FA = false alarms; L = performance index; Q = quality index; K = attention index; **p* < .05.

There was a significant very low correlation coefficient between the scale *negative attitudes towards technology* of the TA-EG (Karrer et al., 2009) and the total number of items attempted by participants ($r = -.20$, $p = .05$), all other correlations were insignificant (all $|r| \leq .188$, all $p \geq .06$). The MANOVA for the effect of device generation on performance indices was insignificant (Pillai's trace = .12, $F_{(12,312)} = 1.11$, $p = .35$). The correlation coefficients between DESIGMA and AST were very low: AST L ($r = .23$, $p = .04$) and AST K ($r = .25$, $p = .02$).

Table 2.

Correlations between test scores of AST, intelligence, and attitudes towards technology

	AST G	AST FV	AST FA	AST L	AST Q	AST K
DESIGMA	.18	-.13	-.15	.23*	.15	.25*
TA 1	-.03	-.03	.03	-.03	.01	-.04
TA 2	.08	.06	.14	.02	-.07	-.02
TA 3	.10	-.11	-.18	.16	.17	.17
TA 4	-.20*	-.16	-.11	-.11	.09	-.12

Notes. G = number of items attempted; FV = misses; FA = false alarms; L = performance index; Q = quality index; K = attention index; TA 1 = enthusiasm; TA 2 = competency; TA 3 = positive attitudes; TA 4 = negative attitudes; **p* < .05.

2.3 Discussion

The aim of study 1 was to assess the feasibility of psychological test development for mobile devices and to evaluate whether the AST is an appropriate computer-based measure for visual sustained attention.

The AST was conducted on students' personal mobile devices without any need for interaction with the researcher, demonstrating that the instructions presented on the screen are sufficient for successful test administration. Moreover, the self-paced nature of the AST allowed individuals to replay the instructions as many times as needed until the task was fully understood. Participants also received immediate feedback during the familiarization period. In classical paper-pencil based diagnostic measures, instructions are rarely repeated, and no individual feedback is presented. However, if test-takers have different levels of task comprehension, resulting test score differences may not only be caused by differences in the underlying trait but also in more shallow confounding variables such as language skills, conscientiousness, or sheer focus on instructions. Thus, in addition to capitalizing on the obvious advantages of mobile devices, the well-known benefits of computer-based testing are inherently available.

Regarding the psychometric properties of the AST the results of the present study demonstrate an excellent split-half reliability coefficient and an excellent model fit to the expected one-factor structure. This is consistent with previous research examining several tests of sustained attention and their intercorrelations, reporting a single factor of sustained attention (Schmidt-Atzert et al., 2006). Regarding the convergent validity, results showed medium to large correlation coefficients between the performance indices of the AST and FAIR-2. This is in accordance with previous research showing comparable correlations between FAIR-2 and d2, or the paper-pencil and computer version of the FAIR-2 (Moosbrugger & Oehlschlägel, 2011). There were no correlations between the error indices of the AST and FAIR-2 which could be explained by very small variance (i.e., participants made only very few mistakes). Attention has been described as prerequisite to many higher cognitive processes (Burack et al., 2012; Krumm et al., 2008; Lezak, 1995; Schumann, 2015). Sternberg (1977) argued that reasoning consists of several processes that also include the (fast) perception of stimulus attributes and relations. While participants need to perceive certain construction rules when solving figural matrices (Becker & Spinath, 2014) they need to search for certain stimulus combinations in the AST to solve the items correctly. This provides a possible cause for the association between the AST and intelligence reported in the present study. Another possible explanation of this association between AST performance and intelligence is that a certain minimum of intelligence might be required to fully understand and solve the task (Moosbrugger & Oehlschlägel, 2011). Importantly, the correlation between the AST and DESIGMA was weaker than between the AST and FAIR-2, further corroborating evidence that the AST primarily measures sustained attention.

Moreover, the convergent validity of the AST is supported by the finding that the test scores were not related to participants' attitudes towards technology. Since previous studies reported that negative attitudes towards technology (e.g., computer anxiety) have an adverse impact on the performance in tests of cognitive ability and educational tests (King et al., 2015; Shirzad & Shirzad, 2017), we examined this issue through the use of the TA-EG (Karrer et al., 2009). We found a small negative correlation coefficient between the subscale "negative attitudes towards technology" and the total number of attempted items in the AST (G). G represents response speed; thus, it is conceivable that the correlation is driven by habitual effects (e.g., people with more negative attitudes towards technology might tend to swipe more slowly than people with less negative attitudes). This notion is supported by findings that emotions (stress, anxiety, depression) of smartphone users can be predicted by their swiping behavior (Balducci et al., 2020). Nonetheless, the absence of associations between the TA-EG and any of the AST performance indices (L, Q, and K) supports the psychometric quality of the AST and indicate that differences in attitudes towards technology do not impair the measurement of sustained attention. Moreover, there was also no effect of device generation on the performance indices, supporting the claim for device independent usability of the AST.

To summarize, study 1 provided evidence for the feasibility of psychological tests designed for administration on mobile devices and it showed that the AST achieved psychometric properties comparable to those of traditional PPTs.

3 Study 2

In study 1, we demonstrated that the AST can be successfully administered in a group setting with a student sample and that its psychometric properties were comparable to traditional test formats. Study 2 aimed to corroborate these findings and investigate whether the AST can also be administered in an unproctored online assessment with a more heterogeneous sample.

3.1 Method

3.1.1 Sample

A total of 256 adults participated. After excluding n = 25 participants for not following the instructions of the AST, the final sample consisted of n = 231 participants (n = 99 female, n = 124 male; n = 8 did not provide information about their gender) with a mean age of 33.22 years (*SD* = 11.14). Eight participants indicated to have dyschromatopsia.

3.1.2 Measurement instruments

In study 2, the AST was administered in exactly the same manner as described in study 1.

3.1.3 Procedure

Participants were recruited on social media (i.e., Facebook, Instagram, & WhatsApp) and received a link to access the study. There were no exclusion criteria for participation, and participants did not receive monetary compensation for their participation. After giving their consent, participants were required to enter sociodemographic information. Subsequently, the AST was administered in a self-paced mode comparable to study 1.

3.1.4 Statistical Analyses

Test and item properties (i.e., item difficulty, part-whole correlation- and reliability coefficients) were analyzed as described in study 1. Furthermore, the association between AST performance and age was analyzed by means of bivariate correlation methods and supported by a test comparing the oldest and youngest quartile with a two-sample *t*-test.

3.2 Results

The psychometric analyses of the AST replicated the findings of study 1. Exact values can be found in the appendix (table A1). There were significant but very low correlation coefficients between age and the G index (r = -.32, p < .001), the L index (r = -.31, p < .001) and the K index (r = -.29, p < .001) with the tendency to a decrease in test performance with increasing age. Furthermore, a comparison of the 25 % oldest and youngest participants revealed a significant difference in the number of items solved ($t_{(114)}$ = -4.03, p < .001, effect size d = 0.75)

3.3 Discussion

Study 2 aimed to replicate the psychometric properties of the AST and extend the findings of study 1 in a more heterogeneous sample within a field setting and without an experimenter's guidance. Results regarding the psychometric properties of the AST replicated the findings reported in study 1. Thus, the AST can be administered on various mobile devices, even in field settings, without impairing its psychometric quality. Accordingly, the AST provides great advantages over traditional paper-pencil tests and CBTs requiring experimenters' guidance in proctored settings. This allows the recruitment of large and heterogeneous samples coming from diverse populations (Tippins, 2015), and including participants from developing countries, who have better access to smartphones than to desktop computers (Joshi & Avasthi, 2007).

Furthermore, performance in the AST was significantly associated with participants' age. This result ties in well with existing research indicating that speeded performance decreases with age (Carriere et al., 2010; Moosbrugger & Oehlschlägel, 2011). The effect was strongest for AST G which represents the response speed but was also found for L (index of attentive performance) and K (index of continuity).

Most importantly though, study 2 provided evidence that psychological achievement tests for mobile devices can be conducted in a self-administered online assessment without any researcher being present to ensure that participants are not distracted from the test.

4 General Discussion

We have argued, that while the advantages of CBTs are widely acknowledged (Tippins, 2015), the possibilities to administer CBTs on mobile devices have yet to be explored more thoroughly. Additionally, test development should focus more on innovative test formats that go beyond *computerized pen-and-paper tests* (Koch, Becker, et al., 2021). However, such aspirations need to be tested for feasibility before broader adaptation. For this purpose, we developed the AST – a test of sustained attention which requires participants to interact intuitively with their smartphones, hence, keeping insecurities with the device minimal. Obviously, the fact that most other tests of sustained attention employ response formats that can be administered on paper as well as computer screens (e.g., crossing something out) is driven by their proven validity and reliability (Blotenberg & Schmidt-Atzert, 2019a; Brickenkamp, 1981, p. 2; Moosbrugger & Oehlschlägel, 2011). Thus, the present study aimed to demonstrate that the AST with its new response format has the necessary psychometric properties for use in psychological diagnostics. In study 1, we found that the AST has excellent psychometric properties, and that performance was not related to the generation of the mobile device, or attitudes towards technology. Furthermore, we found a small association between fluid reasoning and performance in the AST. .Most importantly, correlation coefficients indicating construct validity are comparable to those reported by the authors of the FAIR-2 (Moosbrugger & Oehlschlägel, 2011). Moreover, a CFA with a single factor showed excellent fit indicating that the development was successful. In study 2, there was evidence for the well-established decrease of performance in sustained attention tasks with age (Carriere et al., 2010), and we were able to demonstrate that psychological tests for mobile devices can be self-administered in an unproctored online study. This enables researchers to recruit participants online, and administer the AST, regardless of location and available hardware.

Furthermore, the present study provides evidence the AST test performance was not affected by the generation of the mobile device. This is an important prerequisite if tests are to be administered on test-takers personal mobile devices; an approach to remote testing which has the potential to facilitate psychological testing immensely. One recurring worry of CBT developers and practitioners is rooted in studies reporting differences in test scores depending on attitudes towards technology (King et al., 2015; Shirzad & Shirzad, 2017). We also explored this possibility by administering a measure of attitudes towards technology in study 1, however, we only found a small effect of one subscale (negative attitudes towards technology) on one AST test index. While the present data are certainly not sufficient to draw final conclusions regarding the merits of our intuitive response format, it can be assumed that sorting stimuli by pushing them into different directions may be the type of "usual behavior" individuals show when interacting with mobile devices.

It has been argued that the initial hard- and software costs are the most expensive part of CBTs. The AST was thus developed under the open-source GNU GPLv3 license and can be administered free of charge on test-takers' personal mobile devices. For practitioners, the AST provides a useful addition to their portfolio of testing materials. All required files for using the test as described in this paper can be downloaded from the first author's Github repository (Koch, Möller, et al., 2021) and deployed to any server. For each test-taker, a log of each response as well as response times are provided and can be recoded automatically. If a more secure and standardized version is required, the AST can also be installed on any Android device locally (e.g., a tablet that is used exclusively for diagnostics). The biggest advantage of the AST, however, might lie in its flexibility of administration. Test providers can decide whether participants need to be present in the test lab to take the test (which is recommended for high-stakes testing), or whether they can take the test at home on their own device (which is recommended if the AST is used, for example, as an early screening measure). Furthermore, psychological tests on mobile devices facilitate the diagnostic process with samples who are limited in their mobility (e.g., people with handicap or no access to public transport) as one device is enough to administer a series of measures.

There are two noteworthy limitations to the present study. First, the odd-even split-half reliability often results in an overestimation of the reliability in speeded tests with many items. The aim of the present study was to provide evidence for the feasibility of psychological (achievement) testing on mobile devices and therefore did not include a repeated-measures design. However, with regard to the development and improvement of the AST, further studies should inspect its test-retest reliability with different inter-test intervals. Second, the three-minute version of the AST presented in this study takes only half the time of the FAIR-2 which gives rise to the question whether it really measures sustained attention. While the correlation coefficients between the FAIR-2 and the AST indicate comparability, further studies should explore the information that can be gathered by administering longer versions of the AST. In contrast to traditional paper-pencil tests, this is a strong advantage of the AST, as increasing its duration takes only an adjustment of one value in the test settings.

4.1 Conclusion

To summarize, the present studies provided evidence for the feasibility of psychological tests on (individuals' personal) mobile devices. We decided to test this in the field of sustained attention because it is one of the most important abilities underlying most higher cognitive processes such as working memory or reasoning (e.g., Schumann, 2015). However, a major weakness in existing tests of sustained attention was that no test has been developed specifically for administration on mobile devices. For this purpose, the AST was developed and tested in this study, and was shown to fulfill the necessary psychometric requirements. The AST is a freely available test of sustained attention that can be employed flexibly, and in settings where the administration of PBTs is not feasible. Based on these findings, we strongly advise that future research and test development should explore the potential of mobile devices and to make use of their potential for the measurement of psychological constructs. Tests on mobile devices cannot only be administered more liberally but may also provide a unique opportunity to implement large field studies in a variety of cultural and socioeconomic backgrounds.

5 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Arthur, W., Doverspike, D., Muñoz, G. J., Taylor, J. E., & Carr, A. E. (2014). The Use of Mobile Devices in High-stakes Remotely Delivered Assessments and Testing: Mobile Devices and Remotely Delivered Assessments. *International Journal of Selection and Assessment, 22*(2), 113–123. https://doi.org/10.1111/ijsa.12062

Balducci, F., Impedovo, D., Macchiarulo, N., & Pirlo, G. (2020). Affective states recognition through touch dynamics. *Multimedia Tools and Applications, 79*(47–48), 35909–35926. https://doi.org/10.1007/s11042-020-09146-4

Becker, N., & Spinath, F. M. (2014). *Design a Matrix Test. Ein Distraktorfreier Matrizentest zur Erfassung der Allgemeinen Intelligenz (DESIGMA)* [Measurement instrument]. Hogrefe.

Blotenberg, I., & Schmidt-Atzert, L. (2019a). On the Characteristics of Sustained Attention Test Performance. *European Journal of Psychological Assessment*, 8.

Blotenberg, I., & Schmidt-Atzert, L. (2019b). Towards a Process Model of Sustained Attention Tests. *Journal of Intelligence, 7*(1), 3. https://doi.org/10.3390/jintelligence7010003

Bourdon, B. (1895). Observations comparatives sur la reconnaissance, la discrimination et l'association. *Revue Philosophique de La France et de l'Étranger*, *40*, 153–185. JSTOR.

Brickenkamp, R. (1981). *Test d2 – Aufmerksamkeits-Belastungs-Test (7th ed.)* [Measurement instrument]. Hogrefe.

Brickenkamp, R., Schmidt-Atzert, L., & Liepmann, D. (2010). *Test d2 – Revision (d2-R)* [Measurement instrument]. Hogrefe.

Brown, M. I., & Grossenbacher, M. A. (2017). Can you test me now? Equivalence of GMA tests on mobile and non-mobile devices. *International Journal of Selection and Assessment*, *25*(1), 61–71. https://doi.org/10.1111/ijsa.12160

Burack, J. A., Enns, J. T., & Fox, N. A. (2012). *Cognitive neuroscience, development, and psychopathology. Typical and atypical developmental trajectories of attention* (1st ed.). Oxford University Press.

Büttner, G., & Schmidt-Atzert, L. (Eds.). (2004). *Diagnostik von Konzentration und Aufmerksamkeit*. Hogrefe.

Carriere, J. S. A., Cheyne, J. A., Solman, G. J. F., & Smilek, D. (2010). Age trends for failures of sustained attention. *Psychology and Aging*, *25*(3), 569–574. https://doi.org/10.1037/a0019363

Ghaderi, M., Mogholi, M., & Soori, A. (2015). Comparing Between Computer based Tests and Paper-and-Pencil based Tests. *International Journal of Education and Literacy Studies*, *2*(4). https://doi.org/10.7575/aiac.ijels.v.2n.4p.36

Goldhammer, F., Moosbrugger, H., & Krawietz, S. A. (2009). FACT-2 – The Frankfurt Adaptive Concentration Test: Convergent Validity with Self-Reported Cognitive Failures. *European Journal of Psychological Assessment*, *25*(2), 73–82. https://doi.org/10.1027/1015-5759.25.2.73

Hofer, S. I., Reinhold, F., & Koch, M. (in press). *Students home aloneProfiles of internal and external factors associated with mathematics learning from home. European Journal of Psychology of Education.*

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Joshi, D., & Avasthi, V. (2007). *Position Paper–Mobile Internet UX for Developing Countries*. Mobile HCI 07, Singapore.

Karrer, K., Glaser, C., Clemens, C., & Bruder, C. (2009). Technikaffinität erfassen–der Fragebogen TA-EG. In A. Lichtenstein, C. Stößel, & C. Clemens (Eds.), *Der Mensch im Mittelpunkt technischer Systeme* (pp. 196–201). VDI Verlag GmBH.

King, D. D., Ryan, A. M., Kantrowitz, T., Grelle, D., & Dainis, A. (2015). Mobile Internet Testing: An analysis of equivalence, individual differences, and reactions: Mobile Internet Testing. *International Journal of Selection and Assessment*, *23*(4), 382–394. https://doi.org/10.1111/ijsa.12122

Kingston, N. M. (2008). Comparability of Computer- and Paper-Administered Multiple-Choice Tests for K–12 Populations: A Synthesis. *Applied Measurement in Education*, *22*(1), 22–37. https://doi.org/10.1080/08957340802558326

Koch, M., Becker, N., Spinath, F. M., & Greiff, S. (2021). Assessing intelligence without intelligence tests. Future perspectives. *Intelligence*, *89*. https://doi.org/10.1016/j.intell.2021.101596

Koch, M., Möller, C., & Spinath, F.M. (2021). *Attention Swiping Task* (1.0.0) [JavaScript]. https://doi.org/10.5281/zenodo.5733123

Krumm, S., Schmidt-Atzert, L., & Eschert, S. (2008). Investigating the Structure of Attention: How Do Test Characteristics of Paper-Pencil Sustained Attention Tests Influence Their Relationship with Other Attention Tests? *European Journal of Psychological Assessment, 24*(2), 108–116. https://doi.org/10.1027/1015-5759.24.2.108

Krumm, S., Schmidt-Atzert, L., Schmidt, S., Zenses, E.-M., & Stenzel, N. (2012). Attention Tests in Different Stimulus Presentation Modes: A Facet Model of Performance in Attention Tests. *Journal of Individual Differences, 33*(3), 146–159. https://doi.org/10.1027/1614-0001/a000085

Lezak, M. D. (1995). *Neuropsychological Assessment* (3rd ed.). Oxford University Press.

Matsunaga, M. (2008). Item Parceling in Structural Equation Modeling: A Primer. *Communication Methods and Measures*, *2*(4), 260–293. https://doi.org/10.1080/19312450802458935

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*(3), 449–458. https://doi.org/10.1037/0033-2909.114.3.449

Moosbrugger, H., & Goldhammer, F. (2006). Aufmerksamkeits- und Konzentrationsdiagnostik. In K. Schweizer (Ed.), *Leistung und Leistungsdiagnostik* (pp. 83–102). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-33020-8_6

Moosbrugger, H., & Oehlschlägel, J. (2011). *Frankfurter Aufmerksamkeits-Inventar 2: FAIR-2* [Measurement instrument]. Huber.

Moosbrugger, H., Oehlschlägel, J., & Steinwascher, M. (2011). *Frankfurter Aufmerksamkeits-Inventar 2* (2., überarbeitete, ergänzte und normenaktualisierte Auflage des FAIR von Moosbrugger&Oehlschlägel, 1996). Hogrefe.

Potosky, D., & Bobko, P. (2004). Selection Testing via the Internet: Practical Considerations and Exploratory Empirical Findings*. *Personnel Psychology, 57*(4), 1003–1034. https://doi.org/10.1111/j.1744-6570.2004.00013.x

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. R. (2017). *psych: Procedures for personality and psychological research*.

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., & Yiend, J. (1997). `Oops!’: Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia*, *35*(6), 747–758. https://doi.org/10.1016/S0028-3932(97)00015-8

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2). https://doi.org/10.18637/jss.v048.i02

Schmidt-Atzert, L., Bühner, M., & Enders, P. (2006). Messen Konzentrationstests Konzentration? *Diagnostica*, *52*(1), 33–44. https://doi.org/10.1026/0012-1924.52.1.33

Schumann, F. (2015). *Untersuchung zur prädiktiven Validität von Konzentrationstests—Ein chronometrischer Ansatz zur Überprüfung der Rolle von Itemschwierigkeit, Testlänge und Testdiversifikation*. Technische Universität Chemnitz.

Schweizer, K. (2005). An Overview of Research into the Cognitive Basis of Intelligence. *Journal of Individual Differences*, *26*(1), 43–51. https://doi.org/10.1027/1614-0001.26.1.43

Shirzad, M., & Shirzad, H. (2017). The Effect of Computer Literacy on the Participants' Writing Ability in TOEFL iBT. *Theory and Practice in Language Studies*, *7*(2), 134. https://doi.org/10.17507/tpls.0702.07

Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review*, *84*(4), 353–378. https://doi.org/10.1037/0033-295X.84.4.353

Tippins, N. T. (2015). Technology and Assessment in Selection. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 551–582. https://doi.org/10.1146/annurev-orgpsych-031413-091317

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A Meta-Analysis of Testing Mode Effects in Grade K-12 Mathematics Tests. *Educational and Psychological Measurement*, 20. https://doi.org/10.1177/0013164406288166

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of Computer-Based and Paper-and-Pencil Testing in K–12 Reading Assessments: A Meta-Analysis of Testing Mode Effects. *Educational and Psychological Measurement*, *68*(1), 5–24. https://doi.org/10.1177/0013164407305592

Westhoff, K., & Hagemeister, C. (2005). *Konzentrationsdiagnostik*. Pabst.

Williams, D. C., & Saunders, K. J. (1997). Methodological issues in the study of drug effects on cognitive skills in mental retardation. In N. R. Bray (Ed.), *Int Rev Res Mental Retardation* (pp. 55–109). Academic Press.

Appendix

Table A3.

Mean item difficulties and part-whole correlations for study 2.

Page	Difficulty	Part-Whole Correlation
1	.11	.51
2	.15	.56
3	.14	.60
4	.15	.58
5	.14	.64
6	.15	.64
7	.17	.65
8	.18	.66
9	.20	.67
10	.22	.68
11	.27	.67
12	.34	.64
13	.40	.66
14	.49	.65
15	.56	.62
16	.68	.59
17	.77	.56
18	.84	.53
19	.88	.52
20	.93	.46
21	.95	.41
22	.98	.40
23	.99	.45
24	.99	.46
25	1.00	.44
26	1.00	.44
27	1.00	.44
28	1.00	.44