Psychological Test and Assessment Modeling, 2025.67:141-165

DOI: https://doi.org/10.2440/001-0022

Equivalent Selective Attention Scores From Different Digital Devices: On the Fairness of Assessments Designed for Smartphones, Tablets, and Desktop Computers

Isabell Baldauf[1]*, Maximilian O. Steininger*[2]*, Georg Mandler*[1] *&*
Marco Vetter[1,3,4]

[1] SCHUHFRIED GmbH, Hyrtlstr., 2340 Mödling, Austria
[2] Department of Cognition, Emotion, and Methods in Psychology, University of Vienna
[3] Department of Developmental and Educational Psychology, Faculty of Psychology, University of Vienna, Vienna, Austria
[4] Department of Psychology, University of Salzburg, Salzburg, Austria

Abstract:
Organizations are increasingly offering pre-employment assessments on different digital devices to evaluate candidates. However, in most cases, it remains untested whether the psychometric properties of those assessments are equivalent when different devices are used. Thus, for most assessments, it is unclear whether scores of candidates differing in their choice of device can be compared fairly. The aim of this study is to investigate whether employing a mobile first based cognitive assessment yields equivalent scores of selective attention across different devices. Measurement equivalence across device types was tested using data collected from 296 matched participants. Participants completed the assessment on either a desktop computer or a smartphone. The equivalence of selective attention test scores was investigated using confirmatory multigroup factor analysis. Measurement invariance ensures that test-takers with the same latent trait level have equal probabilities of solving each item and achieving the same scores. Employing a mobile first design approach resulted in equivalent psychometric properties of the assessment in both groups, as indicated by measurement invariance on all levels of investigation. Thus, measurement invariance was achieved unconditionally for both test groups. The results of this study provide convincing evidence that adhering to mobile first principles can yield a valid and reliable assessment for selective attention that can be used on different devices equivalently. The study highlights the importance of considering different devices when designing digital assessments to avoid systematically disadvantaging candidates due to their choice of device.

Keywords: online assessment, mobile testing, mobile first design, cognitive assessment, selective attention

Correspondence: Isabell Baldauf Schuhfried GmbH, Hyrtlstraße 45, 2340 Mödling; baldauf@schuhfried.com

Introduction

Organizations are increasingly utilizing remote psychometric assessments for personnel selection to facilitate accessible and engaging pre-employment evaluations. There are several reasons for the increased interest in conducting assessments online. One reason is the candidate-centric job market, and a necessity for organizations to use the recruitment and hiring process to attract top talent. A way to accomplish this is to offer an efficient and accessible candidate experience using new technologies, such as interactive elements and mobile-delivered assessment (Grelle & Gutierrez, 2019). In an age where a company's online presence is more important than ever, an outdated application process can potentially have a negative impact on the company's image. Another reason for this shift towards remote testing is its simplicity and international accessibility. The COVID-19 pandemic necessitated the development of remote testing solutions since on-site testing was not feasible in a lot of selection scenarios. However, this shift towards remote testing results in applicants taking tests presented as part of an online assessment on their personal devices. Thus, organizations are increasingly lacking control over the types of devices applicants use to complete their assessments.

According to recent statistics, there is a wide variability in personal devices used when accessing the Internet (Eurostat, 2022a). Even in 2016, before the COVID-19 pandemic, cell phones or smartphones were the most common devices and were used by more than three-quarters (79%) of Internet users. This was followed by laptops or notebooks (64%), desktop computers (54%), and tablet computers (44%). By 2021, the proportion of increases remained about the same and even rose to 90% for smartphones (for a detailed comparison see Eurostat (2022b), Eurostat (2022c)). Notably, most digital psychometric assessments are primarily designed for desktop computers, and in many cases, it remains unclear whether administering a test on a device for which it was not specifically designed leads to equivalent test performance across candidates. Furthermore, restricting remote pre-employment assessments to desktop computers poses a challenge for candidates without larger devices at home. These candidates may need to relocate to public spaces offering such devices (e.g., libraries), where they are less likely to complete the tests without interruptions. Thus, designing psychometric assessments solely for desktop computers inevitably introduces issues related to the assessment’s psychometric properties, such as its validity or fairness (Arthur et al., 2014).

Although both companies and candidates are increasingly using a variety of devices during the hiring process, there is limited empirical evidence on the comparability of test scores obtained from these devices. While previous studies have shown little difference between non-cognitive measures (e.g., personality) on mobile and non-mobile devices (Arthur et al., 2014; Ihsan & Furnham, 2018), studies on cognitive measures are inconsistent and scarce. To ensure fairness in the use of assessment results, test developers and test users must be certain that results, whether on item level or overall test scores, are not impacted by variations introduced using different devices. This is the only way to ensure that the respondents do not have any disadvantage because of their choice of device. As highlighted in previous studies (for a review, see Dadey et al., 2018), several factors must be considered when comparing devices, including screen size, displayed content, and the specific item types used in testing.

The issue of screen size, combined with the displayed content, introduces significant challenges to ensuring comparability across devices. One key concern is the method of interaction with the test material. Using a fingertip on a touchscreen instead of a mouse can reduce precision, especially when objects are closely spaced, smaller than the fingertip, or require intricate interactions such as drag-and-drop. Such tasks are more difficult to perform accurately on touchscreens compared to traditional mouse input.

Another challenge arises from the variability in the amount of test content displayed at once, which depends on the screen size. Research suggests that screens of 10 inches or larger are generally suitable for viewing and interacting with test material, with minimal differences in performance at the item level or overall test outcomes (Davis et al., 2013; Keng et al., 2011). However, smaller screens may introduce additional difficulties. For instance, Davis and Strain-Seymour (2013) observed that features such as calculators or on-screen keyboards can obstruct portions of the test content, placing additional demands on participants' working memory. Similarly, Davis et al. (2016) found that students who could not simultaneously view tasks and reading passages reported difficulty retaining the task information while reading. Supporting these findings, Sanchez and Branaghan (2011) demonstrated that small screens, which often require more scrolling, can impede complex cognitive processing. Switching from portrait to landscape mode on smaller screens, however, mitigated some of these negative effects, particularly for participants with lower abilities.

The choice of task or item type also plays a significant role. Eberhart (2015) found that while students generally performed slightly better on desktops than tablets for math and English tests, this effect was task-dependent. Multiple-choice tasks favored desktops, but technology-enhanced items (e.g., selecting text templates) showed no such difference. Davis et al. (2015) observed little impact of device type on tasks requiring simple interactions, such as multiple-choice or hot spot items. Further research by Davis et al. (2016) on seven task types, including drag-and-drop, multiple-choice or hot spot revealed no significant performance differences across devices for most task types. These findings highlight the complex interplay of devices, tasks, and test performance, emphasizing the need for careful test design to ensure device independence.

Mobile first responsive web design

One way to proactively address the issues outlined above is by adhering to mobile-optimized and mobile first responsive web design practices (Grelle & Gutierrez, 2019). In mobile first responsive web design, the developmental process starts with the smallest supported device and works up to larger devices to provide a user experience that is optimized for and consistent across all device types (Ward, 2017). First applications of this principle show promising results for certain cognitive abilities. For example, mobile-optimized assessments for measures of working memory (Frost et al., 2018; Morgan et al., 2018) or general cognitive tests (Gutierrez & Grelle, 2018; Schuhfried, 2024) have been shown to be equivalent across device types after applying mobile first principles. These studies emphasize that optimizing item design for the smallest screen size makes it necessary to revise classic item formats. Given the available display space, classic cognitive ability tasks designed for paper-pencil or desktop applications often feature elaborate graphics or text. For example, tests of reasoning or reading comprehension typically present a section of text or information in tables or charts, and candidates must read or review this information to select an answer option or answer a question (Sanchez & Goolsbee, 2010). Simply reducing the size of these tasks would bias the scores because the tasks would no longer be clearly visible. Thus, new paradigms and test concepts need to be developed. Although past studies have already shown, that adhering to mobile first responsive web design when constructing device-independent assessments leads to promising results for certain cognitive abilities, the range of assessments is still restricted to a handful of cognitive domains. The aim of the current study is to expand the range of tests available by designing a device-independent test for measuring selective attention.

Selective Attention

Selective attention is a core component of human information processing that enables individuals to selectively process relevant stimuli while simultaneously inhibiting irrelevant or competing information (Duncan & Humphreys, 1989; Lavie, 2005). It serves as a filtering mechanism within the broader framework of attentional control, allowing the organism to allocate limited cognitive resources efficiently in environments characterized by high perceptual load (Lavie, 2005). Selective attention differs from other attention constructs such as sustained attention — which involves maintaining focus over prolonged periods — or divided attention, which refers to the simultaneous monitoring of multiple stimuli or tasks. In contrast, selective attention is defined by the dynamic prioritization of task-relevant input and the suppression of interference from distractors.

In this study, selective attention is operationalized as the ability to rapidly and accurately discriminate between highly similar visual stimuli based on abstract rule-based criteria. The task paradigm requires participants to identify target configurations within a stimulus array under strict time constraints, thereby engaging both stimulus-driven and goal-directed attentional processes. This construct is particularly relevant in applied domains where individuals must make accurate perceptual discriminations in the presence of noise or competing information (e.g., driving, aviation, monitoring tasks).

Measuring selective attention is a critical component of many assessment scenarios, as an individual's ability to concentrate is a prerequisite for performing a wide range of tasks. Concentrated work is essential for performance in all activities that require conscious perception and information processing (Westhoff & Hagemeister, 2005). This fundamental ability is considered a relevant predictor for efficient or safe behavior in various professions, such as pilots (Hunter & Burke, 1994), drivers (Vetter et al., 2018), train drivers (Guo et al., 2019), or workers (Goertz et al., 2014). Low scores of the skill or permanent (e.g., due to illness) and temporary impairments (e.g., due to impairing substances) may have serious consequences (Arthur et al., 1994; Baysari et al., 2008). Due to its elementary character, attention is relevant to nearly all practical and intellectual activities (Sturm, 2009). Thus, investigating attention is essential for a large variety of psychological assessment scenarios. Traditional test paradigms for selective attention, such as those used in the FAIR-2 (Moosbrugger & Oehlschlägel, 2011), COG (Schuhfried, 2019) or d2-R (Brickenkamp et al., 2010), involve identifying and marking or connecting symbols that are visually similar and difficult to discriminate. A closer examination of these paradigms reveals that a 1:1 transfer of paper-pencil or desktop tests to a smartphone is not feasible. Small icons, crowded items, and the requirement to mark and connect symbols with a pencil make it exceedingly challenging to ensure comparability across different devices.

Main aims of the present study

As outlined above, selective attention plays a critical role in various assessment contexts, yet existing test paradigms often lack adaptability across digital devices. To address this limitation, we develop a novel test based on mobile-first design principles. While previous research demonstrates the potential of mobile-first approaches in cognitive ability assessment (Gutierrez & Grelle, 2018; Morgan et al., 2018), their application to selective attention remains unexplored.

The present study pursues two main objectives. First, we aim to develop and validate a selective attention test that adheres to established psychometric quality standards while functioning equivalently across devices. Second, we examine whether mobile-first design supports measurement equivalence by comparing test performance between desktop and smartphone administration. To this end, we employ a confirmatory multigroup factor analysis to evaluate configural, metric, and scalar invariance, ensuring that individuals with equivalent latent trait levels demonstrate consistent item responses and total scores regardless of device.

Method

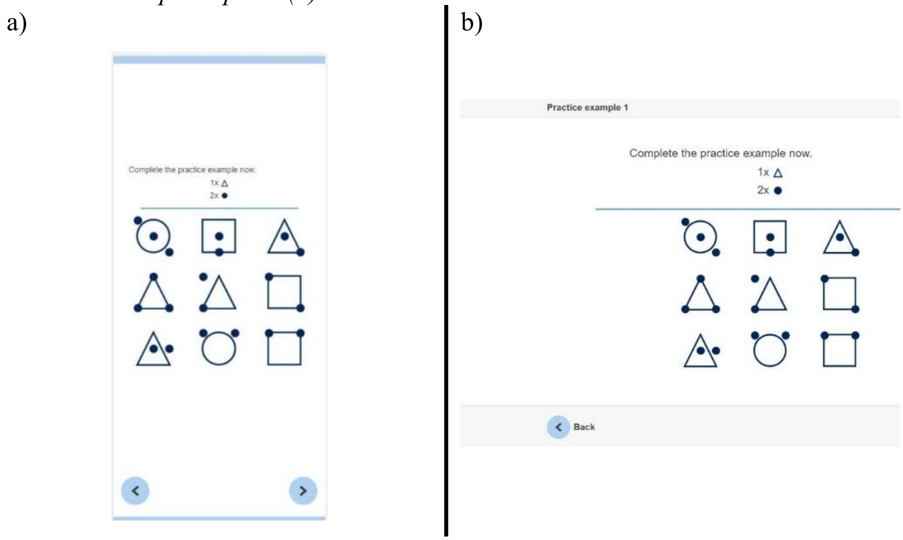
Materials

To create a device-equivalent measure of selective attention, item development focused on two main areas: ensuring comparability across different devices and constructing the item material based on theory. First, when designing the test, we put a special emphasis on deriving the item content from cognitive theories on selective attention. Thus, a major aim of designing the assessment was to ensure content and construct validity and a suitable distribution of the item difficulty (Embretson, 1998). To this end, we created a cognitive model, outlining the skills and knowledge required to process the item material and providing a framework to empirically quantify the impact of these factors on correct task performance (Arendasy & Sommer, 2011; Gorin, 2006). This approach facilitated the assessment of construct and content validity at the level of individual items (Gorin, 2006).

Second, by adopting the mobile-first concept from the field of web design (Ward, 2017) and adhering to recommendations regarding the comparability of test results on different platforms (Dadey et al., 2018), we aimed to ensure outcome equivalence across devices. Thus, during development, it was crucial that the type of device used for the assessment did not affect the likelihood of a correct response, and that there were no overall performance differences between the devices. To this end, particular attention was paid to the paradigm used and the display and presentation of the item content. In adherence to mobile first principles, we optimized the content in the first step for use on a smartphone and then transferred it to larger devices. This ensured that the content could be depicted in a comparable manner and that all relevant information was available on all devices. As part of regular functionality reviews (see Way et al., 2016) during development, the items were checked for whether they could be displayed similarly, i.e. that the position or the relative size of the symbols and selection elements was comparable and that they were not displayed distortedly. In addition, insights from existing studies were considered when designing the items. For example, it was ensured that all content of the items was available at the same time and in all orientations of the device (horizontal and vertical) to minimize the impact of working memory (e.g., due to scrolling or popups; Sanchez & Branaghan, 2011). Furthermore, the presented scope of information was kept constant across all devices (Winter, 2010). Finally, we ensured that the input boxes and the selection of the answers were sufficiently sized on all devices and avoided certain answer types (e.g., free text input) that could not be implemented in comparable form on all devices.

Following these principles, we developed a selective attention test. In this test, the respondent's task was to identify and mark stimuli, within a 3x3 grid, that matched specific target criteria, while simultaneously ignoring stimuli that did not match these criteria (see Figure 1). Each individual stimulus consisted of an unfilled basic shape (triangle, circle, or square) and one or multiple filled detail shapes (triangle, circle, or square). Target stimuli at the top of the page specified which of the nine stimuli below must be marked for each item. These target stimuli also defined the required basic shape and the specific number and type of detailed shapes that needed to appear in a stimulus for it to be marked. For example, if the item contained a large, unfilled triangle and two small, filled circles, all stimuli in the 3x3 grid that exactly matched this configuration had to be marked. A time limit of 6 seconds per item was set, which was in line with both the time limit suggested by Moosbrugger and Oehlschlägel (2011) and the general requirement for unambiguous interpretation (Schmidt-Atzert et al., 2004), as the working time per item was kept constant for all respondents, thus ensuring comparability of accuracy.

Figure 1
Depiction of a single item of the selective attention test (TACO) on a smartphone (a) and on a desktop computer (b)


Complete the practice example now.

Complete the practice example now.

The stimuli and structure of the items were selected according to the recommendations for constructing attention tests (Moosbrugger & Oehlschlägel, 2011; Westhoff & Hagemeister, 2005), as well as structuring device-independent psychometric tests. First, the test included items requiring the discrimination of relatively simple stimuli that were clearly perceivable. Second, we selected easily discriminable and culture-neutral shapes. Third, the stimuli varied across three dimensions: two dimensions (i.e., the basic shape and the detailed shape) were relevant to the solution, while the third dimension (i.e., the position or arrangement of the detailed shapes), was irrelevant. Fourth, the stimuli had to be combined according to a rule that was easy to remember, minimizing the potential confounding influence of memory. Fifth, the target stimulus switched per test page and thus throughout the test. For the selection of the target stimuli, attention was paid to maintaining a balance of the symbols used.

To create items with different degrees of complexity, the search stimuli were varied in their difficulty. Based on information processing theories and empirical findings, four predictors were considered: the density of information per stimulus and test page (i.e., the number of individual shapes, including basic and detailed shapes), the number of target stimuli to be marked, the similarity between distractors and target stimuli, and the spatial proximity of the target stimuli. Following previous robust findings, we assumed that a higher density of information per test page or individual stimulus required more serial processing resources (Forster & Lavie, 2007; Hyman, 1953; Jensen & Munro, 1979). Thus, we expected lower probabilities of correctly completing an item as the amount of information presented increased. In addition, we aimed to vary the difficulty by increasing the number of target stimuli to be marked, as we assumed that verifying a correct stimulus would require more cognitive resources than falsifying an incorrect stimulus. Furthermore, we expected that a high degree of similarity between distractors and target stimuli would increase the difficulty of an item (Becker, 2011; Duncan & Humphreys, 1989; Lavie, 2005; Pashler, 1987). Finally, we assumed that a higher spatial proximity between the target stimuli would favor faster processing. To operationalize selective attention the number of correctly worked individual stimuli was defined as the main outcome variable. It consists of all correctly marked target stimuli and all correctly ignored distractors, and describes the ability to maintain selective attention for a short period (5 minutes) at a high level.

Participants

The data was collected in 2019-2020 in the test and research center of SCHUHFRIED GmbH according to a representative, randomized, and parallelized design. The participants were assigned to one of the two groups (desktop computer or smartphone) using stratified randomization. Both groups reflected a representative distribution of the German-speaking population and were matched in terms of age (in 5-year categories), gender, and education level. A total of 148 (30.8%) participants from the sample completed the test on a smartphone while the other 331 (69.2%) participants completed the test on a computer via the mouse. The sample consisted of 255 (53.2%) women and 224 (46.8%) men aged 14 to 90 ($\bar{x}$ = 48.64; *SD* = 18.81). In total, 3 (0.6%) participants did not have any school-leaving qualification (EU education level 1), 34 (7.1%) participants completed mandatory school or secondary modern school, but without completed vocational training (EU education level 2), 155 (32.4%) participants had a completed vocational training or a qualification from a college (EU education level 3), 182 (38%) participants completed a higher school with general qualification for university entrance (EU education level 4) and 105 (21.9%) participants had a university degree (EU education level 5). To test for measurement invariance, a subset of the sample was parallelized. In total, the sample for the measurement invariance testing contained 296 participants. The sample was evenly split between computer and smartphone users (50% each), with a gender distribution of 46% male and 54% female, an average age of 48.19 years (*SD* = 19.6), and educational levels ranging from EU Level 1 (1%) to Level 5 (20%), with the majority at Levels 3 (32%) and 4 (43%). To check whether the collected sample was representative of the population of interest, a *chi*^{2} test considering participants' gender and age was performed (*chi*^{2}[29]=24.29, *p*=.71).

Procedure

All participants completed a standardized test sequence consisting of five computerized assessments:

1. the new Selective Attention Test,
2. the Trail-Making Test – Langensteinbach Version S1 (TMT-L; Rodewald et al., 2019),
3. the Simultaneous Capacity/Multitasking Test S2 (SIMKAP; Bratfisch & Haman, 2018),
4. the N-Back Verbal Test S4 (NBV; Schellig & Schuri, 2019), and
5. the Cognitrone S2 (COG; Schuhfried, 2019).

The entire testing session lasted approximately 60 minutes. The order of the tests was fixed across all participants. All tests were administered in a supervised setting using the Vienna Test System (VTS) software platform. Participants were randomly assigned to one of two experimental conditions, differing only in the device used to complete the Selective Attention Test: either a desktop computer or a smartphone. The remaining tests in the sequence were completed on a desktop device for all participants.

To reflect the increasingly fluid boundaries between device categories (e.g., smartphones, tablets, laptops), we categorized devices based on screen resolution. Devices with a screen resolution greater than 767×1023 pixels were classified as desktop computers, while devices with screen resolutions equal to or below this threshold were classified as smartphones. Smartphone content was optimized using a mobile-first design approach, including responsive scaling and touch-based input, while the desktop version maintained a fixed layout optimized for mouse and keyboard interaction. However, the core test content — including stimuli, timing, and scoring — remained identical across platforms.

All participants received standardized on-screen instructions and completed one practice trial before beginning the actual task. The Selective Attention Test consisted of 30 trials, each comprising a 3×3 matrix of stimuli with a target template displayed at the top. Each matrix was shown for 3.5 seconds followed by a 1.5-second response interval. This structure was identical for both device groups.

Statistical Analyses

To examine device independence, a two-step analysis procedure was employed. First, several quality indicators were assessed, including the plausibility of the construction rationale, factorial and construct validity, and internal consistency of the test. These analyses provided a foundation for evaluating the overall quality of the test and its suitability for further comparison across devices. In the second step, measurement invariance analysis was performed to assess the comparability of the test outcome (Vandenberg & Lance, 2000) between desktop and smartphone conditions. This ensured that the test outcomes were equivalent across device types, confirming that the selective attention test was independent of the device used.

For the empirical investigation of the construction rationale, a multiple linear regression analysis was conducted using the full representative sample ($n = 479$). Four factors influencing item difficulty — information density, number of target stimuli, similarity between distractor and target stimuli, and spatial proximity of target stimuli — were selected as predictor variables. These factors were incorporated into the regression model using the backward method, with the mean number of correctly solved symbols per test page (i.e., symbols correctly marked or ignored) as the dependent variable.

To evaluate convergent and discriminant validity for the selective attention test, we constructed a nomological network and tested it within the Cattell-Horn-Carroll (CHC) model framework (Schneider & McGrew, 2018). Tests included in the analysis had to meet four criteria: (1) sufficient reliability (≥ .7) to avoid artificial correlation attenuation (Lance et al., 2006), (2) symmetry in generality to ensure adequate comparisons, (3) inclusion of only ability tests to avoid validity restrictions due to differing measurement methods (Bühner, 2011), and (4) representative samples with diverse sociodemographic distributions (age, gender, education) to avoid sampling bias. Based on these criteria, we selected the following tests: TMT-L Version S1 (Rodewald et al., 2019), SIMKAP S2 (Bratfisch & Haman, 2018), NBV S4 (Schellig & Schuri, 2019), and COG S2 (Schuhfried, 2019). Construct validity was assessed by correlating the main outcome of the selective attention test with external indicators, including both convergent and discriminant tests. Convergent validity was demonstrated through correlations with TMT-L and COG tests, as all three tests align with the CHC secondary factors "Reaction & decision speed" (Gt) or "Processing Speed" (Gs) under the primary factor "General cognitive speed."

Additionally, we inspected the factor structure of the selective attention test using confirmatory factor analysis (CFA), which included all 50 test pages. To ensure reliable estimation, the items were grouped into ten parcels, maximizing systematic variance and reducing random response effects (Little et al., 2002). Each item parcel represented the sum of correctly solved symbols across grouped items, and these values were incorporated into the model. The MLR estimator from the R package "lavaan" (R Core Team, 2022; Rosseel, 2012) was used to estimate the model (Reinecke, 2014). Model fit was evaluated using indices *CFI* (≥ 0.90), *SRMR* (≤ 0.08), and *RMSEA* (<

0.08), which indicate a strong correlation between the empirical covariance matrix and the theoretical factor model (Hu & Bentler, 1999; Yu, 2002).

Test reliability was assessed using Cronbach's alpha for both the full representative sample and subsamples for each device. Reliability values above .7, .8, and .9 were considered adequate, good, and excellent, respectively (EFPA, 2013). Additionally, descriptive statistics, including means, standard deviations, test statistics, and effect sizes, are presented for each device condition. Cohen's d_p, calculated using the pooled standard deviation of both groups as the denominator, was used to determine effect sizes, ensuring consistency and comparability between the device groups (Jané et al., 2024).

Lastly, we performed a measurement invariance analysis to assess the comparability of selective attention test scores between smartphone and desktop presentations using confirmatory multigroup factor analysis (CFA; Milfont & Fischer, 2015). Following this approach, the dataset was divided into groups (e.g., smartphone vs. desktop), and the model fit was assessed for each group separately before conducting multigroup comparisons. This allowed us to examine whether respondents interpreted test items similarly across devices (Bialosiewicz et al., 2013). The four stages of measurement invariance — configural, metric, scalar, and strict — were tested. Configural invariance tested whether the overall factor structure was similar across groups. Metric invariance examined whether factor loadings were equivalent, enabling valid comparisons of factor variances and covariances. Scalar invariance tested whether item intercepts were consistent across groups, allowing comparisons of factor means. Finally, strict invariance, which ensures equal unexplained variance for each item across groups, indicates identical measurement at the item level, though it is often considered too strict to achieve (Deshon, 2004; Lubke et al., 2003).

Results

To investigate the construction rationale, we conducted a multiple linear regression and included four independent variables into the model to predict the mean number of correctly solved symbols. We observed that all predictors, except for the similarity of the target stimuli and the distractors (p = .45), as well as spatial proximity of the target stimuli (p = .09) showed significant impacts on the outcome variable (see Table 2). Thus, following the backward method of independent variable integration, the final model included four variables - the density of information per stimulus and test page (corresponds to the number of individual shapes, i.e., the basic or detailed shapes), the number of target stimuli to be marked, the similarity between distractors and target stimuli, as well as the spatial proximity of the target stimuli to be marked - which collectively explained 84.1% of the variance in the outcome (R = .92, adjusted R^2 = .83, F = 81.08, p < .01, df=3).

Table 1
Regression model between the previously defined predictor variables of the items and their mean number of correctly worked symbols.

			95% *CI*				
Predictor	*B*	*SE*	*LL*	*UL*	*ß*	*T*	*p*
Step 1							
Intercept	10.430	0.459	9.507	11.355		22.738	<.001
Number of target stimuli	-0.364	0.065	-0.496	-0.233	-0.50	-5.596	<.001
Density of information	-0.042	0.008	-0.057	-0.027	-0.48	-5.637	<.001
Similarity distractor/ target stimulus	0.019	0.025	-0.032	0.070	0.06	0.763	.450
Spatial proximity target stimuli	-0.118	0.067	-0.253	0.018	-0.11	-1.743	.088
Step 2							
Intercept	10.697	0.298	10.098	11.296		35.941	<.001
Number of target stimuli	-0.385	0.059	-0.504	-0.267	-0.53	-6.554	<.001
Density of information	-0.044	0.007	-0.058	-0.030	-0.50	-6.151	<.001
Spatial proximity target stimuli	-0.133	0.067	-0.248	0.021	-0.10	-1.692	.097

***Note*: R^2 = .843 for step 1 (*p* < .01), ΔR^2 = .002 for step 2 *(p* = .45).**

The number of target stimuli [*beta* = -0.385, *SE* = 0.059, *t*= -6.554, *p* = <.001 two-tailed, 95% *CI* = [-0.504, -0.267]] and the density of information [*beta* = -0.044, *SE* = 0.007*, t*= -6.151, *p* = <.001 two-tailed, 95% *CI* = [-0.058, -0.030]] showed a similar standardized beta that was considerably higher than the influence of the spatial proximity of the target stimuli. The latter variable showed a non-significant effect at the conventional alpha level (β = -0.133, SE = 0.067, *t* = -1.692, *p* = .097, two-tailed, 95% *CI* = [-0.248, 0.021]), but the direction of the effect was consistent with theoretical expectations. In summary, we found that the three predictor variables explained a significant proportion of the variance in test performance. This confirms that most of the theoretically relevant factors considered in item design were empirically supported, validating the successful construction of the item material based on the construction rationale.

Construct validity was assessed by correlating the main outcome of the test with external indicators, including both construct-related (convergent) and construct-unrelated (discriminant) tests.

Table 2
Correlations between the primary outcome of the selective attention test (TACO) and convergent or discriminant tests

Test	1	2	3	4	5	6
1. TACO						
2. TMT – BTA	-.694					
3. TMT – BTB	-.716	.693				
4. COG	-.630	.439	.597			
5. SIMKAP – SIM	.700	-.645	-.655	-.532		
6. SIMKAP – STQ	.075	-.149	-.145	.041	.262	
7. NBV	.271	-.210	-.306	-.263	.349	.105

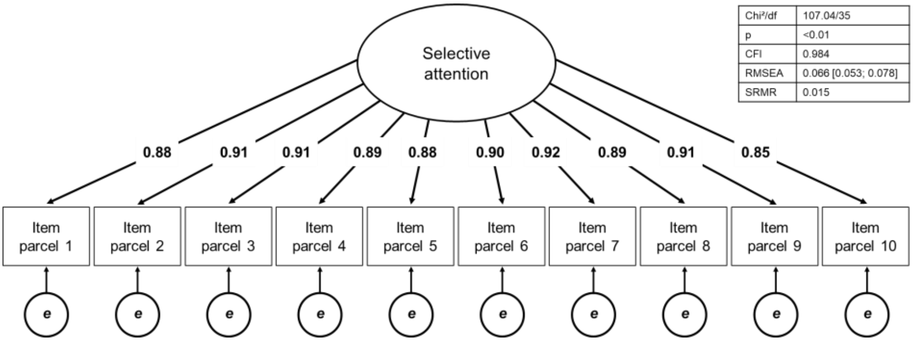
Note: TMT = Trail-Making-Test; COG= Cognitrone; SIMKAP = Simultaneous Capacity/Multitasking; NBV = NBack Verbal

In line with our predefined nomological network we observed high correlations (see Table 2) of the selective attention test with TMT-L and COG which can be considered as good and adequate (EFPA, 2013). Divergent validity was tested by correlating the selective attention test with the SIMKAP, and NBV tests. The SIMKAP is most clearly associated with the primary factor Processing speed within the CHC model. Consistent with our prediction, the selective attention test showed a high correlation with the variable Simultaneous capacity (r = .7) and a weak correlation with the variable Stress tolerance (r = .075). As expected, no correlation with the variable Stress tolerance was shown in the SIMKAP. Lastly, as expected, we observed a low correlation between the performance in the selective attention test and the NBV test (r = .271). The NBV test measures verbal working memory and is associated with the primary factor Working memory (Gwm) in the CHC model. In summary, we confirmed construct validity through correlations consistent with the expectations derived from our nomological network.

After assessing construct validity, we examined factorial validity by analyzing the factor structure of the selective attention test through confirmatory factor analysis (see

Figure 2). All test items (i.e., *k* = 50 test pages) were included in the analysis and grouped into ten item parcels. Figure 2 presents the factor structure, factor loadings and performance indicators of the model. The theoretically proposed model demonstrated a very good fit to the data [$\chi^2(35) = 107.039$; *CFI* = 0.984; *RMSEA* = 0.066 (90% *CI*: 0.053; 0.078); *SRMR* = 0.015] confirming a good factorial validity of the test with high loadings (≥ .85) for each item parcel.

Figure 2
Confirmatory factor analysis of the mean number of correctly worked symbols



Chi²/df	107.04/35
p	<0.01
CFI	0.984
RMSEA	0.066 [0.053; 0.078]
SRMR	0.015

In a final step, we used data from 296 matched participants to empirically investigate measurement equivalence across device types. Descriptive statistics for all primary and secondary variables of the test were calculated for each group and are presented in Table 3.

Table 3

Descriptive statistics for the primary and secondary variables of the selective attention test, presented separately by device group

Variables	Device	*n*	$\bar{x}$	*SD*	Test statistic	Effect size	Cronbach's alpha
Selective attention	Desktop	148	387.3	39.7	$t(294) = 0.196$, $p = .845$	$d = 0.02$	$\alpha = .976$
	Smartphone	148	388.3	55.4			$\alpha = .972$
Number of omissions	Desktop	148	56.7	35.5	$t(294) = -0.319$, $p = .750$	$d = 0.04$	$\alpha = .976$
	Smartphone	148	55.4	36.2			$\alpha = .977$
Number of false alarms	Desktop	148	6.4	12.3	$t(294) = 0.280$, $p = .779$	$d = 0.03$	$\alpha = .951$
	Smartphone	148	6.0	11.3			$\alpha = .945$

Note: *n* = sample size, $\bar{x}$ = sample mean, *SD* = standard deviation, *d* = Cohen's *d*

First, we performed a two-sampled *t*-test comparing both groups to determine if the sample means in the outcome variables were equal. We observed no significant mean differences in test scores across different devices, with effect sizes (Cohens' *d*) ranging from $d = 0.02$ to $d = 0.04$, and thus well below the conventional threshold ($d = 0.2$) for small effects (Cohen, 1988). Second, we estimated Cronbach's alpha separately for each subgroup and across subgroups. We observed high internal consistency for each of the variables in the full sample ($\alpha = .973$, .976, and .938), as well as in the subgroups using desktop computers ($\alpha = .976$, .976, and .951) or smartphones ($\alpha = .972$, .977, and .945). Notably, the reliability estimates for each subgroup showed substantial overlap, indicating comparable levels of internal consistency across groups. Third, we used a measurement invariance analysis to investigate the equivalence of the selective attention test scores on different devices (Byrne, 2001; Kline, 1998). We tested measurement invariance between the two groups for all levels of invariance (i.e. metric, scalar, or strict). We used changes in $\chi 2$ values, CFI, and RMSEA for the nested models as indicators of measurement invariance (Putnick & Bornstein, 2016). Following this approach, measurement invariance was indicated if there was no significant change in the $\chi 2$ value (Byrne et al., 1989), and if changes in the *CFI* and *RMSEA* did not exceed value of $|0.01|$ and $|0.015|$, respectively (Rutkowski & Svetina, 2014; Chen, 2007). The results for the measurement invariance analysis for selective attention compared for smartphone versus desktop presentation are presented in Table 4.

Table 4

Measurement invariance analysis comparing selective attention performance for smartphone and desktop presentation

Model	χ^2	*df*	*CFI*	*RSMEA*	$\Delta\chi^2$	*p*	*ΔCFI*	*ΔRSMEA*
Smartphone	55.539	35	0.987	0.063				
Desktop	65.533	35	0.980	0.077				
Configural invariance	141.08	70	0.984	0.070	--	--	--	--
Metric invariance	154.90	79	0.983	0.067	10.938	.28	0.001	0.003
Scalar invariance	165.65	88	0.982	0.065	10.777	.29	0.001	0.002
Strict invariance	180.44	98	0.983	0.060	9.820	.45	0.001	0.005

Note: df = degrees of freedom, *CFI* = comparative fit index, *RMSEA* = root mean square error of approximation

The factor models showed good model fits for both subgroups (see Table 4). Importantly, we observed no significant changes in the *χ2* value. Furthermore, changes in both the *CFI* or *RMSEA* for successive measurement models were well below the thresholds of |0.01| and |0.015|, respectively. Notably, this was true even at the most stringent level of strict invariance. Thus, measurement invariance can be assumed unconditionally over both test groups (smartphone vs. desktop), indicating the successful development of a test that is equivalent across devices.

Discussion

The primary aim of this study was to develop a device-independent test for measuring selective attention. Given the relevance of selective attention in many assessment contexts and the general lack of device-independent measures for this ability, we sought to address the growing demand for fair assessments across different platforms by constructing a test that is equivalently applicable on desktop computers and smartphones. To this end, we developed a new test paradigm grounded in mobile first principles and informed by cognitive theories. We assessed quality indicators of the test – including the construction rationale, internal consistency, construct validity, and factorial validity – and examined device independence by comparing test performance on desktop computers and smartphones, while also investigating measurement invariance across both device types.

We found that the test performed well across all quality indicators. First, we demonstrated that the hypothesized predictor variables for item difficulty significantly accounted for the observed variance in the main outcome variable. Second, we obtained excellent internal consistency estimates in both the full sample and the subsamples, further supported by a factor model that showed a good fit to the data and high factor loadings for all item parcels. Third, we found evidence for both convergent and divergent validity, with correlational patterns aligning with expectations for tests assessing related and unrelated ability domains. Additionally, we showed that test performance was comparable across participants using smartphones or desktop computers. Importantly, we also demonstrated measurement invariance across both devices at all levels of invariance, including the most stringent form of strict invariance. Thus, by applying mobile first principles and grounding item development in cognitive theories, we successfully created a test for selective attention that can be fairly and comparably used on both smartphones and desktop computers.

Demonstrating that adherence to these principles results in device-equivalent measurement is important, as it shows that tests can be successfully designed and implemented across various devices. This not only facilitates fair comparisons but also offers flexibility in test administration, which is often desired by both companies and test takers. Simply transferring the same measurement to other devices without prior evidence of equivalence may severely disadvantage certain test takers, resulting in unfair and inefficient assessment outcomes. While it has been successfully demonstrated that non-cognitive (Arthur et al., 2014; Ihsan & Furnham, 2018) and cognitive measures (Frost et al., 2018; Gutierrez & Grelle, 2018; Morgan et al., 2018) can be constructed by following these principles, evidence regarding selective attention is scarce. Notably, expanding this evidence to include a measure of selective attention is not only practically relevant but also underscores that a broad range of measurements can be developed without discriminating against participants based on their choice of device.

To show that the newly developed test was truly comparable across different devices, a confirmatory multigroup factor analysis was conducted on a parallelized sample of 296 participants. Measurement invariance was tested in four stages (configural, metric, scalar and strict or residual invariance), all of which were well supported by the data. Configural invariance showed that the relationships between each item in the test and the latent factor it was intended to measure, were consistent across both groups. Achieving metric invariance demonstrated that differences in factor variances and covariances were not due to group-based differences in the properties of the test itself. Scalar invariance assessed whether the item intercepts were equivalent across groups, and the multigroup model fit indicated no significant mean differences between the groups tested. Lastly, strict invariance revealed that the unexplained variance for each item was equal across groups, indicating identical measurement at the item level of the construct for both groups. Notably, the result of strict invariance testing should be interpreted with caution, as we parceled the items rather than integrating them individually in the model.

In addition to observing measurement invariance, we found no evidence of performance differences or discrepancies in internal consistency across parallelized groups further supporting equivalent measurement. When directly comparing test performance across all three outcome variables, we observed non-significant results with effect sizes close to zero, indicating that the difference was not only statistically insignificant but also practically negligible. The observed mean difference of 387.3 and 388.3 in the main variable corresponds to a 1% change in percentile rank within the representative norm sample of the test, which is negligible in most assessment scenarios. Additionally, both subgroups were characterized by excellent Cronbach's alpha values across all three outcome variables. Together with the configural invariance findings from the multigroup factor analysis this provides converging evidence, that the internal consistency of the test was not only of high quality but importantly comparable across groups. In addition to internal consistency, we also observed promising results regarding other quality indicators such as factorial, construct, and content validity.

Regarding content validity, we conducted a multiple linear regression analysis to test the quality of the construction rationale. In the full sample of 479 participants, we found that three predictors (i.e., information density per stimulus and test page, the number of target stimuli to be marked, and the spatial proximity of the target stimuli to be marked) explained 84.1% of the variance in the number of solved symbols per test page. This substantial explanatory power supports the successful construction of the item material, confirming that nearly all the theoretical predictors we hypothesized based on previous literature (Becker, 2011; Duncan & Humphreys, 1989; Forster & Lavie, 2007; Hyman, 1953; Jensen & Munro, 1979; Lavie, 2005; Pashler, 1987; Moosbrugger & Oehlschlägel, 2011; Westhoff & Hagemeister, 2005) were also empirically relevant. Notably, we did not find a significant influence of the similarity between distractors and target stimuli, suggesting that participants could discriminate distractors from targets more easily than expected.

Regarding construct validity we observed adequate to good (EFPA, 2013) correlations with both construct-related tests (i.e., TMT and COG). Additionally, the correlations with construct-unrelated tests were in the expected direction and magnitude for the SIMKAP and NBV. Lastly, we examined the factor structure of the selective attention test using confirmatory factor analysis. We observed a sufficient correlation between the empirical covariance matrix and the covariance matrix expected based on the theoretically postulated factor model. Thus, in addition to the promising findings for content and construct validity we also found evidence for the test's factorial validity.

In summary, the results of this study provide convincing evidence that a construct-oriented, mobile first design can yield a valid and reliable test applicable across various devices. Our goal was not only to develop a measure that showed equivalence across platforms, but also to expand the range of device-independent tests by incorporating a cognitive domain highly relevant to job performance. This is important, as selective attention is a fundamental skill essential for numerous activities requiring conscious perception and processing of information (Westhoff & Hagemeister, 2005). Its elementary character makes attention relevant to nearly all practical and intellectual activities (Arthur et al., 1994; Baysari et al., 2008; Guo et al., 2019; Goertz et al., 2014; Hunter & Burke, 1994; Vetter et al., 2018). Although the demand for device-independent tests is substantial - and such tests are already used in the hiring processes of several companies - empirical studies on the comparability of cognitive ability assessments across devices remains scarce (e.g. Gutierrez & Grelle, 2018; Morgan et al., 2018). Given society's increasing reliance on smartphones, addressing this gap is vital for developing fair and psychometrically sound assessments tailored to mobile platforms. Our findings underscore that simply transferring a traditional test paradigm designed for paper-pencil or desktop use directly to mobile devices is insufficient. Instead adopting a mobile first approach to test development ensures that quality indicators are met while simultaneously enabling device-independent measurement equivalence, ultimately leading to fairer testing practices. Bridging this gap in the future is critical to establishing equitable assessment practices across devices, ensuring fairness for test takers and meeting the demands of modern, technology-driven societies.

Limitation and Future Directions

This study demonstrates the successful implementation of a mobile-enabled cognitive test with a device independent design. While we aimed to accommodate a wide range of device types and sizes, it was not feasible to encompass all variations, particularly given the rapid pace of market innovations (e.g., foldable phones, VR/AR glasses). Consequently, future studies are needed to evaluate whether test performance remains comparable across a wider range of technical implementations. Moreover, the test environment in our study was standardized, with all participants - whether using a smartphone or a desktop computer – completing the test under controlled conditions in the laboratory. This standardization is crucial to control environmental factors (e.g., brightness and noise) across groups, ensuring a comparison with high experimental control and minimizing the influence of potentially confounding variables. However, future studies should not only investigate whether the choice of device might systematically disadvantage test takers but also explore how environmental conditions, which may differ based on the device used, could impact test performance (e.g. Hygge & Knez, 2001; Realyvásquez-Vargas et al.2020).

Given the sustained interest in mobile-enabled cognitive testing within the field of personnel selection, there are many opportunities for additional research. A key future goal should be to develop mobile first design tests that span various cognitive ability domains, enabling the assessment of a broad spectrum of cognitive skills directly on smartphone. This would not only offer a more user-friendly approach to cognitive tests for personnel selection but could also expand the scope of clinical testing applications. Finally, as the emphasis on creating engaging, innovative, and brand-enhancing tests for personnel selection grows, future research should explore the extent to which mobile first test designs are perceived as more interesting by candidates and whether they enhance the overall candidate experience. This could provide valuable insights into how test formats impact candidate engagement and satisfaction, potentially influencing recruitment outcomes and employer branding.

CONFLICT OF INTEREST

Three of the authors Isabell Baldauf, Georg Mandler and Marco Vetter are currently employed by SCHUHFRIED GmbH. The fourth author Maximilian O. Steiniger was also employed by SCHUHFRIED GmbH in the past.

Literature

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Arendasy, M. E. & Sommer, M. (2011). Automatisierte Itemgenerierung: Aktuelle Ansätze, Anwendungen und Forschungen. In L.F. Hornke, M. Amelang, M. Kersting, N. Birbaumer, D. Frey, J. Kuhl et al. (Hrsg.), *Enzyklopädie für Psychologie: Methoden der Psychologischen Diagnostik* (S. 215–280). Göttingen: Hogrefe Verlag.

Arthur Jr, 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. *International Journal of Selection and Assessment*, 22(2), 113-123.

Arthur, W., Keiser, N. L. & Doverspike, D. (2017). An information-processing-based conceptual framework of the effects of unproctored internet-based testing devices on scores on employment-related assessments and tests. *Human Performance*, *31*(1), 1–32. Routledge.

Arthur, W. J., Strong, M. H. & Williamson, J. (1994). Validation of a visual attention test as a predictor of driving accident involvement. *Journal of Occupational and Organizational Psychology*, *67*, 173–182.

Baysari, M. T., McIntosh, A. S. & Wilson, J. R. (2008). Understanding the human factors contribution to railway accidents and incidents in Australia. *Accident Analysis and Prevention*, *40*(5), 1750–1757.

Becker, S. I. (2011). Determinants of dwell time in visual search: Similarity or perceptual difficulty? *PLoS ONE*, *6*(3), 1–5.

Becker, N., Preckel, F., Karbach, J., Raffel, N. & Spinath, F. M. (2014). Die Matrizenkonstruktionsaufgabe. *Diagnostica*, *1*, 22–33.

Bialosiewicz, S., Murphy, K. & Berry, T. (2013). *An introduction to measurement invariance testing: Resource packet for participants*. Retrieved from http://comm.eval.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=63758fed-a490-43f2-8862-2de0217a08b8

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson.

Bratfisch, O. & Haman, E. (2018). *Manual Simultankapazität/Multi Tasking (Version 53)*. Mödling: Schuhfried.

Brickenkamp, R., Schmidt-Atzert, L., & Liepmann, D. (2010). *Test d2-Revision: Aufmerksamkeits-und Konzentrationstest*. Göttingen: Hogrefe.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, application and programming.* London: Lawrence Erlbaum.

Byrne, B. M., Shavelson, R. J. & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). New York, NY: Routledge.

Dadey, N., Lyons, S. & Depascale, C. (2018). The comparability of scores from different digital devices : A literature review and synthesis with recommendations for practice. *Applied Measurement in Education*, *31*(1), 30–50. Routledge.

Dadey, N., Lyons, S. & Depascale, C. (2018). The comparability of scores from different digital devices : A literature review and synthesis with recommendations for practice. *Applied Measurement in Education*, *31* (1), 30–50. Routledge.

Davis, L. L., Janiszewska, I., Schwartz, R., & Holland, L. (2016, March). *NAPLAN device effects study*. Melbourne, Australia: Pearson

Davis, L. L., Orr, A., Kong, X., & Lin, C. (2015). Assessing student writing on tablets. *Educational Assessment*, 20, 180– 198. doi:10.1080/10627197.2015.1061426

Davis, L. L., & Strain-Seymour, E. (2013). *Keyboard interactions for tablet assessments*. Washington, DC: Pearson Education. Retrieved from

Chen, F. F. (2007). *Sensitivity of goodness of fit indexes to lack of measurement invariance.* Structural Equation Modeling, 14(3), 464–504. https://doi.org/10.1080/10705510701301834

Davis, L. L., Strain-Seymour, E., & Gay, H. (2013). *Testing on tablets: Part II of a series of usability studies on the use of tablets for K–12 assessment programs.* Washington, DC: Pearson Education. Retrieved from http://researchnetwork.pearson.com/wp-content/uploads/Testing-on-Tablets-Part-II_formatted.pdf

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, 46(1), 137–149

Duncan, J. & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*(3), 433–458.

Eberhart, T. (2015). *A comparison of multiple-choice and technology-enhanced item types administered on computer versus iPad* (Doctoral dissertation), University of Kansas, Lawrence, KS.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3* (3), 380–396. American Psychological Association.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S.H. Irvine & P.C. Kyllonen (Hrsg.), *Item generation for test development* (S. 219–250). Mahwah. NJ: Lawrence Erlbaum.

European Federation of Psychologists‘ Associations. (2013). *Performance requirements, context definitions and knowledge & skill specifications for the three EFPA levels of qualifications in psychological assessment*. www.efpa.eu/download/5cb1dbc81322f855b6aaf3f96be6a61a (Accessed on 28 March 2019)

Eurostat (2022a). *Internet use by individuals: Almost 8 out of 10 internet users in the EU surfed via a mobile or smart phone in 2016.* https://ec.europa.eu/eurostat/documents/2995521/7771139/9-20122016-BP-EN.pdf/f023d81a-dce2-4959-93e3-8cc7082b6edd (Accessed on 10 October 2022)

Eurostat (2022b). *Households - availability of computers.* https://ec.europa.eu/eurostat/databrowser/view/isoc_ci_cm_h/default/bar?lang=en (Accessed on 10 October 2022)

Eurostat (2022c). *Trust, security and privacy - smartphones (2020 onwards).* https://ec.europa.eu/eurostat/databrowser/view/ISOC_CISCI_SP20__custom_3529383/default/bar?lang=en (Accessed on 10 October 2022)

Forster, S. & Lavie, N. (2007). High perceptual load makes everybody equal: Eliminating individual differences in distractibility with load. *Psychological Science*, *18*(5), 377–381.

Frost, C., Carpenter, J., & Ferrell, J. (2018, April). *Demonstrating equivalence of high-fidelity cognitive measures on mobile devices.* In A. S. Boyce and S. Gutierrez (Chairs), Mobile first design: The key to effective mobile cognitive testing? Symposium conducted at the 33th annual Society for Industrial and Organizational Psychology Conference, Chicago, IL.

Gorin, J. S. (2006). Test Design with Cognition in Mind. *Educational Measurement: Issues and Practice*, *25* (4), 21–35.

Guo, M., Hu, L. & Ye, L. (2019). Cognition and driving safety: How does the high-speed railway drivers’ cognitive ability affect safety performance? *Transportation Research Part F: Traffic Psychology and Behaviour*, *65*, 10–22. Elsevier Ltd.

Gutierrez, S. L. & Grelle, D. (2018). *Impact of mobile-first design on equivalence for cognitive tests.* In S. L. Gutierrez & A. S. Boyce (Chairs), Mobile first design: The key to effective mobile cognitive testing? Symposium presented at the 33rd Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Grelle, D. & Gutierrez, S. (2019). Developing device-equivalent and effective measures of complex thinking with an information processing framework and mobile first design principles. *Personnel Assessment and Decisions*, *5*(3), 21–32

Hornke, L. F., Etzel, S. & Rettig, K. (2019). *Manual Adaptiver Matrizentest (Version 51 - Revision 1)*. Mödling: Schuhfried.

Hornke, L. F., Küppers, M. & Etzel, S. (2000). Konstruktion und Evaluation eines adaptiven Matrizentests. *Diagnostica*, *46* (4), 182–188.

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.

Hunter, D. R. & Burke, E. F. (1994). Prediciting aircraft pilot-training success. *The International Journal of Aviation Psychology*, *4*(4), 297–313.

Hygge, S., & Knez, I. (2001). Effects of noise, heat and indoor lighting on cognitive performance and self-reported affect. Journal of environmental psychology, 21(3), 291-299.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, *45*(3), 188–196.

Ihsan, Z., & Furnham, A. (2018). The new technologies in personality assessment: A review. *Consulting Psychology Journal: Practice and Research*, 70(2), 147.

Jané, M. B., Xiao, Q., Yeung, S. K., Ben-Shachar, M. S., Caldwell, A. R., Cousineau, D., ... & Feldman, G. (2024). Guide to effect sizes and confidence intervals. DOI: https://doi.org/10.17605/OSF. IO/D8C4G.

Jensen, A. R. & Munro, E. (1979). Reaction time, movement time, and intelligence. *Intelligence*, *3*(2), 121–126.

Keng, L., Kong, X. J., & Bleil, B. (2011, April). *Does size matter? A study on the use of netbooks in K–12 assessment.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford Press.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say?. *Organizational research methods*, *9*(2), 202-220.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences*, *9*(2), 75–82.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. Structural equation modeling, 9(2), 151-173.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance. *British Journal of Mathematical and Statistical Psychology*, 56(2), 231–248. https://doi.org/10.1348/000711003770480020

Milfont, T. L., & Fischer, R. (2015). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, 3, 111–130. doi:10.21500/20112084.857

Moosbrugger, H. & Oehlschlägel, J. (2011). *Frankfurter Aufmerksamkeits-Inventar 2: FAIR-2*. Bern: Huber.

Morgan, K. E., LaPort, K. A., Lowery, B. S., Cottrell, J. M., Rangel, B., Martin, N. R., & Boyce, A. S. (2018). *The quest for equivalence: Mobile-first working memory assessment*. S.L. Gutierrez and A. S. Boyce (Chairs), Mobile first design: The key to effective mobile cognitive testing? Symposium presented at the 33rd Annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Pashler, H. (1987). Target-distractor discriminability in visual search. *Perception & Psychophysics*, *41*(4), 285–292.

Putnick, D. L. & Bornstein, M. H. (2016). Measurement invariance conventions and reporting. *Developmental Review*, *41*(9), 71–90.

R Core Team. (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Verfügbar unter: https://www.r-project.org/

Realyvásquez-Vargas, A., Maldonado-Macías, A. A., Arredondo-Soto, K. C., Baez-Lopez, Y., Carrillo-Gutiérrez, T., & Hernández-Escobedo, G. (2020). The impact of environmental factors on academic performance of university students taking online classes during the COVID-19 Pandemic in Mexico. Sustainability, 12(21), 9194.

Reinecke, J. (2014). *Strukturgleichungsmodelle in den Sozialwissenschaften* (2. Aufl.). München: Oldenbourg.

Rodewald, K., Weisbrod, M. & Aschenbrenner, S. (2019). *Manual Trail Making Test (Version 53 - Revision 1)*. Schuhfried.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Rutkowski, L. & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31–57.

Sanchez, C. A. & Branaghan, R. J. (2011). Turning to learn: Screen orientation and reasoning with small devices. *Computers in Human Behavior*, *27* (2), 793–797. Elsevier Ltd.

Sanchez, C. A., & Goolsbee, J. Z. (2010). Character size and reading to remember from small displays. Computers & Education, 55(3), 1056-1062.

Schellig, D. & Schuri, U. (2019). *Manual N-Back Verbal (Version 26 - Revision 1)*. Mödling: Schuhfried.

Schneider, W. J. & McGrew, K. S. (2018). The Cattell-Horn-Carroll theory of cognitive abilities. In D.P. Flanagan & E.M. McDonough (Hrsg.), *Contemporary intellectual assessment: Theories, tests, and issues* (4. Aufl., S. 73–163). New York, NY: Guilford Press.

Schuhfried, G. (2019). *Manual Cognitrone (COG; Version 52 – Revision 2).* Schuhfried.

Schuhfried (2022). *Manual Aufmerksamkeits- und Konzentrationstest (TACO; Version 52 - Revision 2).* Schuhfried GmbH.

Schuhfried (2024). *Manual Inventar zur Testung kognitiver Fähigkeiten (INT; Version 55 - Revision 3).* Schuhfried GmbH.

Sturm, W. (2009). Aufmerksamkeitsstörungen. In W. Sturm, M. Herrmann & T.F. Münte (Hrsg.), *Lehrbuch der klinischen Neuropsychologie* (S. 375–379). Würzburg: Spektrum Akademischer Verlag.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational research methods, 3(1), 4-70.

Vetter, M., Schünemann, A. L., Brieber, D., Debelak, R., Gatscha, M., Grünsteidel, F. et al. (2018). Cognitive and personality determinants of safe driving performance in professional drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *52*, 191–201.

Ward, C. (2017). *Jump start responsive web design*, *2nd edition*. Melbourne, Victoria, Australia: SitePoint Pty. Ltd.

Way, W. D., Davis, L. L., Keng, L. & Strain-Seymour, E. (2016). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In F. Drasgow (Hrsg.), *Technology in testing: Improving educational and psychological measurement* (2. Aufl., S. 260–284). Abingdon, UK: Routledge.

Westhoff, K. & Hagemeister, C. (2005). *Konzentrationsdiagnostik*. Lengerich: Pabst Science Publishers.

Winter, P. (2010). Comparability and test variations. In P. Winter (Hrsg.), *Evaluating the comparability of scores from achievement test variations* (S. 1–11). Washington, DC: Council of Chief State School Officers.

Yu, C. Y. (2002). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes*. (Doctoral dissertation). University of California.