The Web-Based Assessment of Mental Speed:

An Experimental Study of Testing Mode Effects for the Trail-Making Test

Timo Gnambs

Leibniz Institute for Educational Trajectories, Germany

Article type: Brief report

Word count: 2,541

Author Note

Timo Gnambs https://orcid.org/0000-0002-6984-1276

I have no conflicts of interest to disclose. The study was not preregistered. The data, material, computer code, and analysis results are available at https://osf.io/qh4z9/.

Correspondence concerning this article should be addressed to Timo Gnambs, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamber, Germany, E-mail: timo.gnambs@lifbi.de.

Accepted for publication in the *European Journal of Psychological Assessment*.

Abstract

Although web-based cognitive assessments have gained increasing attention in recent decades, it is still debated whether unstandardized test settings allow for comparable measurements as compared to proctored testing, particularly for speeded cognitive tests. Therefore, two within-subject experiments (N = 73 and 72) compared differences in means, criterion correlations with measures of intelligence, and subjective test quality perceptions of a trail-making test between a proctored paper-based, a proctored computerized, and an unproctored web-based administration mode. The results in both samples showed equivalent means between the two computerized modes, equivalent criterion correlations between the three modes, and no differential item functioning. However, the web-based tests were rated as having an inferior measurement quality as compared to the proctored assessments. Thus, web-based testing allows for comparable measurements of mental speed as compared to traditional computerized tests, although it is still regarded as a lower quality medium by test takers.

Keywords: processing speed; mode effect; computerized testing; web-based testing; equivalence test

The Web-Based Assessment of Mental Speed:

An Experimental Study of Testing Mode Effects for the Trail-Making Test

Mental speed is a basic component of current models of intelligence and refers to the ability to efficiently solve simple tasks. Popular indicators of mental speed are trail-making tests (TMT) that require respondents to sequentially connect randomly arranged numbers or letters on a sheet of paper or computer screen as quickly as possible (Oswald, 2016). Because of their robust associations with fluid and crystallized intelligence (Sheppard & Vernon, 2008), these tests are routinely used for neuropsychological screening to quickly gauge a respondent's basic cognitive status. While TMTs have been traditionally administered as paper-based tests (PBT), computer-based test formats (CBT) allow for a stronger standardization and more reliable measurements. However, comparative analyses showed that computerizations of TMTs also produce pronounced mode effects (e.g., Drapeau et al., 2007; Fellows et al., 2017). Different versions of the TMT not only show substantial mean-level differences, but also limited convergent validities. A reason for the incomparability of TMT scores might be that the motor skills required to draw on a sheet of paper are quite different from using a mouse and clicking on a computer screen. However, even tablet administrations with touch functionality do not seem to alleviate this problem (Bracken et al., 2019). In recent decades, cognitive assessments are also increasingly administered over the Internet without the presence of a supervisor (e.g., Steger et al., 2020; Zinn et al., 2021). While web-based tests (WBT) allow similar presentation modes as CBT, the test setting is significantly less standardized, for example, concerning environmental distractors (e.g., noise, lighting conditions). Consequently, unproctored WBT adds another layer of complexity to the assessment process. So far, speeded tests are not routinely administered as WBTs.

Therefore, the aim of the current study was the comparison of TMTs administered as WBT to traditional PBT and CBT versions. Two within-subject experiments evaluated whether mean scores and criterion correlations with measures of intelligence are equivalent between different assessment modes. Because of the similar presentation formats, WBT was

expected to yield comparable scores to CBT, but lower test performance as compared to PBT (cf. Bracken et al., 2019; Drapeau et al., 2007).

Method

Sample Size Rationale and Participants

The study aimed at identifying differences between assessment modes of at least a quarter of a standard deviation (i.e., Cohen's $d = 0.25$). Because short-term retest correlations for the TMT typically fall between .85 and .95 (e.g., Oswald, 2016; Wagner et al., 2011), correlations between TMTs administered in different modes of at least .75 were expected. This resulted in a minimal relevant effect size of Cohen's *dz* of 0.35. An *a priori* power analysis suggested a required sample size of 65 to identify significant differences between two paired samples with a power of .80 and a significance level of $\alpha = .05$. A respective equivalence test of paired means required a sample size of at least 69. The study included two independent samples with Sample 1 comprising $N = 73$ students (45 women) from a medium-sized university in Austria and Sample 2 including $N = 72$ students (57 women) from an Austrian upper secondary school and a mid-sized university. Consequently, the first sample ($M = 28.70$, $SD = 7.28$) was significantly older (Cohen's $d = 1.57$, 95% CI [1. 20, 1.94]) than the second ($M = 19.50$, $SD = 3.85$). In both samples, most participants (about 90%) were native speakers of German and right-handed. None of the participants were excluded because of visual or motor impairments.

Instruments

Mental speed was measured with a TMT (*Zahlen-Verbindungs-Test*; Oswald, 2016) that required respondents to connect circled numbers from 1 to 90. The numbers were arranged in a matrix structure with 9 rows and 10 columns on a single page. The test includes four tasks (A to D) that are administered successively and are combined to a total mean score. Following Oswald (2016), two different administration and scoring procedures were adopted in the two samples. In Sample 1 the test was administered individually and respondents were instructed to connect all 90 numbers as quickly as possible without setting a time limit. The

time required to finish the task (in seconds) represented the respective score. In contrast, in Sample 2 the test was administered in small groups of up to 10 participants. After a time limit of 30 seconds respondent scores were determined as the number of correctly connected numbers. Otherwise, the test procedure was identical in both samples

Fluid intelligence was measured with 30 items from the *Advanced Progressive Matrices* (Raven, Raven, & Court, 1998) that required respondents to identify logical rules to complete a depicted pattern. The test was administered on a computer with a time limit of 30 minutes. The omega reliabilities in the two samples were $\omega = .84$, and $\omega = .93$, respectively. Crystallized intelligence was measured with 20 items from a vocabulary test (*Mehrfachwahl-Wortschatz-Intelligenztest*; Lehrl, 2005) that required respondents to identify real words from a set of four pseudowords. The test was administered on a computer without a time limit. The omega reliabilities in the two samples were $\omega = .84$ and $\omega = .75$, respectively.

The perceived measurement quality and face validity of the different TMT versions was captured with two subscales from the *AKZEPT* questionnaire (Kersting, 2008) with four items each on six-point response scales from 1 = “does not apply” to 6 = “applies completely”. Two additional subscales were administered but not further analyzed (see supplemental material). For the subscale measurement quality, the omega reliabilities in the two samples ranged from $\omega = .74$ to .79 and from $\omega = .69$ to .80, respectively. The respective reliabilities for the subscale face validity ranged from $\omega = .75$ to .82 and from $\omega = .76$ to .84.

Experimental Procedure

The experiments adopted one-factorial within-subject designs with three conditions that represented the administration mode of the TMT (PBT, CBT, WBT). The PBT and CBT were administered in the lab, while the WBT was administered within one week of the proctored assessment by inviting the participants by email to finish the web-based version of the TMT. Participation in the WBT required the use of a laptop or personal computer. The WBT could not be accessed using mobile devices with small screens. In each mode, the participants received a written instruction and were administered two practice tasks that

presented smaller versions of the TMT including 20 numbers to guarantee that the participants correctly understood the test instruction. Then, the four items of the TMT and the self-report questionnaire were presented. After finishing the two TMTs in the lab, the respondents were also administered proctored and computerized tests of fluid and crystallized intelligence. The sequence of the three testing modes was partially randomized. Further details on the procedure and the computerized test are given in the supplemental material.

Analysis Plan

The assessment modes were compared by testing for equivalence of means (Mara & Cribbie, 2012) using an effect of Cohen's *dz* of ±0.35 as equivalence bounds and for equivalence of criterion correlations (Counsell & Cribbie, 2015) using $r = .20$ as a minimal relevant effect. The significance level for all analyses was set to $\alpha = .05$. Differential item functioning (DIF) for the four TMT tasks was examined using generalized linear mixed-effects regressions (see Van den Noortgate & De Boeck, 2005). In this approach, item difficulty parameters are represented by fixed effects, person abilities are given by random effects, and DIF is indicated by cross-level interactions between the item effects and two dummy-coded grouping variables representing the mode (using WBT as reference category).

Results

The TMTs administered in the two samples exhibited comparable omega reliabilities of .95 / .92 for the PBT, .96 / .90 for the CBT, and .96 / .96 for the WBT. Moreover, all three versions exhibited similar measurement structures as indicated by the regression models with DIF that did not outperform a model only acknowledging main effects but no DIF, $\chi^2(df = 6) = 3.37$, $p = .761$ for Sample 1 and $\chi^2(df = 6) = 1.90$, $p = .929$ for Sample 2. In addition, the size of the estimated DIF effects fell between Cohen's $d = -0.08$ and 0.17 in Sample 1 and between Cohen's $d = -0.03$ and 0.22 in Sample 2 and did not reach our threshold for non-negligible effects (see the supplemental material for full results).

The different versions of the TMT were substantially correlated. In Sample 1, the WBT administration correlated with the PBT and CBT administration at $r(73) = .76$ and $r(73)$

= .85, while the respective correlations were $r(72) = .85$ and $r(72) = .82$ in Sample 2. Thus, the web-based administration did not systematically impair the convergent validities as compared to the correlations of $r(73) = .86$ and $r(72) = .80$ that were observed in the two proctored assessments. However, the mean structure exhibited systematic differences between modes (see Table 1). The participants showed significantly ($p < .05$) inferior test performances on paper as compared to a computer or web-based administrations in both samples. The respective effect sizes (Cohen's *d*) fell at .25 and .34 in Sample 1 and at -0.44 and -0.48 in Sample 2. In contrast, the equivalence tests substantiated comparable means for the two computerized administrations (Cohen's $d = 0.09$ and 0.05). Finally, for the criterion correlations of the three TMTs with measures of fluid and crystallized intelligence the equivalence tests supported comparable criterion correlations for most mode comparisons (see Table 1). The largest difference of $\Delta r = .12$ was observed in Sample 1 between PBT and CBT for the correlations with reasoning scores which resulted in ambiguous inference tests that did not allow definite conclusions. In contrast, the differences in criterion correlations for the remaining comparisons were smaller ($\Delta r \leq .08$) and corroborated equivalence.

The respondents also evaluated the perceived test quality of the TMTs administered in the different modes. The measurement quality of the WBT was rated significantly inferior as compared to the other assessment modes, while the significant equivalence tests summarized in Table 1 corroborated comparable user perceptions between PBT and CBT. However, the respective effect sizes showing a disadvantage for the WBT were only slightly larger than our threshold for non-negligible effects and fell between Cohen's $d = 0.25$ and 0.34 in the two samples. In contrast, the perceived face validity was rated highest for CBT with mean differences reaching up to Cohen's $d = 0.57$, while the equivalence tests showed comparable ratings of face validity for PBT and WBT.

Discussion

Testing mode effects in mental speed frequently result in nonequivalence between paper-based and computer-based administrations (e.g., Bracken et al., 2019; Drapeau et al.,

2007; Fellows et al., 2017), thus making meaningful comparisons difficult. Although one could argue that the advantages of computerized assessments outweigh the need for psychometric equivalence, mode effects impede the implementation of mixed-mode designs and can even misdirect theory development if paper- and computer-based measures of a construct seem different as a result of testing modes alone (see Schmitz & Wilhelm, 2019). With the proliferation of web-based cognitive testing, the test setting might also contribute to the incomparability of test scores. Therefore, the present study compared in two within-subject experiments unproctored web-based TMTs to their proctored counterparts. On the one hand, the results replicated previous findings on the nonequivalence of means for PBT, despite comparable measurement models across modes. On the other hand, CBT and WBT led to comparable results indicating that WBT does not introduce unique heterogeneity in the measurement of mental speed. Importantly, the criterion validities with intelligence were comparable between modes which might serve as an initial indicator that comparable constructs were measured. However, perceptions of test quality by the test takers suggested that WBT is still viewed as an inferior assessment mode, despite its promising psychometric characteristics. In summary, these findings support the feasibility of measuring mental speed using web-based versions of the TMT, as long as mean-level comparisons with PBT are not of primary interest. It might even be feasible to apply CBT norms to web-based TMTs. Thus, WBT could increase the wider accessibility of psychological testing and allow reaching more diverse samples for which traditional proctored testing might be difficult. However, these findings must be viewed as preliminary. The highly selective samples do not allow generalizations to more heterogeneous populations such as patients with cognitive impairments or older age groups. Lower computer literacy might lead to more pronounced difficulties with CBT or WBT. Moreover, the examined nomological net was limited to two measures of intelligence and, thus, need to be extended in future studies to other relevant constructs such as working memory or occupational performance.

References

Bracken, M. R., Mazur-Mosiewicz, A., & Glazek, K. (2018). Trail Making Test: Comparison of paper-and-pencil and electronic versions. *Applied Neuropsychology: Adult, 26*, 522-532. https://doi.org/10.1080/23279095.2018.1460371

Counsell, A., & Cribbie, R. A. (2015). Equivalence tests for comparing correlation and regression coefficients. *British Journal of Mathematical and Statistical Psychology, 68*(2), 292-309. https://doi.org/10.1111/bmsp.12045

Drapeau, C. E., Bastien-Toniazzo, M., Rous, C., & Carlier, M. (2007). Nonequivalence of computerized and paper-and-pencil versions of Trail Making Test. *Perceptual and Motor Skills, 104*(3), 785-791. https://doi.org/10.2466/pms.104.3.785-791

Fellows, R. P., Dahmen, J., Cook, D., & Schmitter-Edgecombe, M. (2017). Multicomponent analysis of a digital Trail Making Test. *Clinical Neuropsychologist, 31*(1), 154-167. https://doi.org/10.1080/13854046.2016.1238510

Kersting, M. (2008). Zur Akzeptanz von Intelligenz- und Leistungstests [On the acceptance of intelligence and aptitude tests]. *Report Psychologie, 33*, 420-433.

Lehrl, S. (2005). *Mehrfachwahl-Wortschatz-Intelligenztest* [Multiple-Choice Vocabulary Intelligence Test]. Hogrefe.

Mara, C. A., & Cribbie, R. A. (2012). Paired-samples tests of equivalence. *Communications in Statistics - Simulation and Computation, 41*(10), 1928-1943. https://doi.org/10.1080/03610918.2011.626545

Oswald, W. D. (2016). *Zahlen-Verbindungs-Test* [Trail-Making Test]. Hogrefe.

Schmitz, F., & Wilhelm, O. (2019). Mene mene tekel upharsin: Clerical speed and elementary cognitive speed are different by virtue of test mode only. *Journal of Intelligence, 7*:16. https://doi.org/10.3390/jintelligence7030016

Sheppard, L. D., & Vernon, P.A. (2008) Intelligence and speed of information-processing: A review of 50 years of research. *Personality and Individual Differences, 44*(3), 535-551. https://doi.org/10.1016/j.paid.2007.09.015

Steger, D., Schroeders, U., & Gnambs, T. (2020). A meta-analysis of test scores in proctored and unproctored ability assessments. *European Journal of Psychological Assessment, 36*, 174-184. https://doi.org/10.1027/1015-5759/a000494

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics, 30*(4), 443-464. https://doi.org/10.3102/10769986030004443

Wagner, S., Helmreich, I., Dahmen, N., Lieb, K., & Tadić, A. (2011). Reliability of three alternate forms of the trail making tests A and B. *Archives of Clinical Neuropsychology, 26*(4), 314-321. https://doi.org/10.1093/arclin/acr024

Zinn, S., Landrock, U., & Gnambs, T. (2021). Web-based and mixed-mode cognitive large-scale assessments in higher education: An evaluation of selection bias, measurement bias, and prediction bias. *Behavior Research Methods, 53*, 1202-1217. https://doi.org/10.3758/s13428-020-01480-7

Table 1

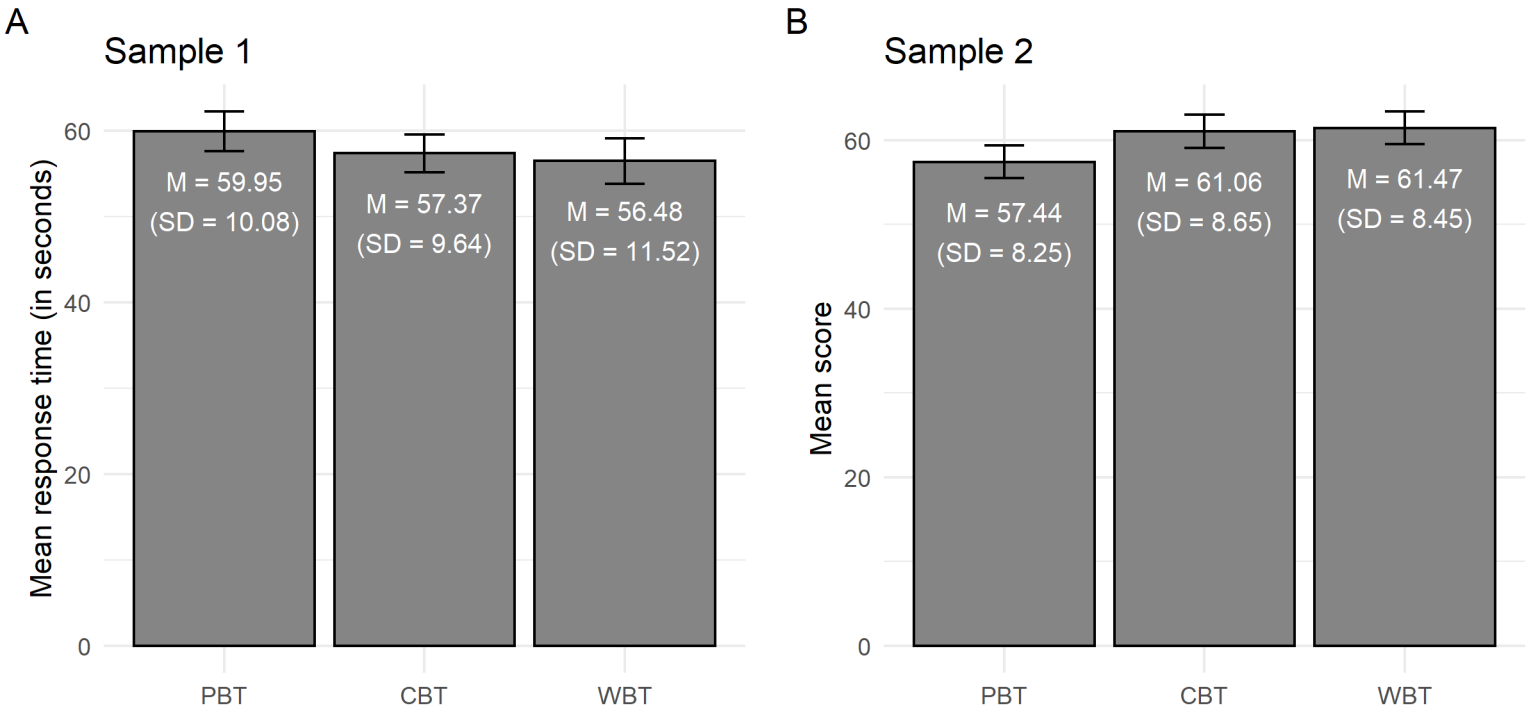
Summary of Nil Hypotheses and Equivalence Tests for Mode Effects

	Sample 1					Sample 2				
	Effect size (95% CI)	$t_{nil}(df)$	p_{nil}	$t_{equ}(df)$	p_{equ}	Effect size (95% CI)	$t_{nil}(df)$	p_{nil}	$t_{equ}(df)$	p_{equ}
Trail-Making Test										
PBT vs. CBT	0.25 (0.12, 0.38)	4.15 (72)	<.001	1.12 (72)	.868	-0.43 (-0.60, -0.27)	-5.72 (71)	<.001	2.72 (71)	.996
PBT vs. WBT	0.34 (0.17, 0.51)	3.88 (72)	<.001	0.85 (72)	.802	-0.48 (-0.63, -0.34)	-7.54 (71)	<.001	-4.54 (71)	1.00
CBT vs. WBT	0.09 (-0.03, 0.22)	1.26 (72)	.213	-1.77 (72)	.041	-0.05 (-0.19, 0.09)	-0.66 (71)	.510	2.34 (71)	.011
Measurement Quality										
PBT vs. CBT	0.11 (-0.09, 0.30)	1.16 (72)	.252	-1.87 (72)	.033	-0.07 (-0.21, 0.06)	-1.10 (71)	.274	1.90 (71)	.031
PBT vs. WBT	0.32 (0.10, 0.54)	3.02 (72)	.003	-0.00 (72)	.499	0.25 (0.04, 0.46)	2.32 (71)	.023	-0.68 (71)	.248
CBT vs. WBT	0.25 (0.06, 0.45)	2.39 (72)	.020	-0.64 (72)	.262	0.34 (0.12, 0.56)	3.04 (71)	.003	0.37 (71)	.515
Face Validity										
PBT vs. CBT	-0.57 (-0.79, 0.36)	-5.76 (72)	<.001	-2.73 (72)	.996	-0.16 (-0.32, 0.01)	-1.75 (71)	.084	1.25 (71)	.108
PBT vs. WBT	0.02 (-0.16, 0.19)	0.17 (72)	.862	-2.85 (72)	.003	0.05 (-0.11, 0.20)	0.60 (71)	.554	-2.41 (71)	.009
CBT vs. WBT	0.57 (0.36, 0.79)	5.76 (72)	<.001	2.74 (72)	.996	0.18 (0.00, 0.37)	2.06 (71)	.044	-0.95 (71)	.173
Fluid intelligence										
PBT vs. CBT	.12 (.02, .22)	1.97 (70)	.053	[a]	.096	-.01 (-.13, .11)	-0.11 (69)	.909	[a]	.002
PBT vs. WBT	.03 (-.10, .16)	0.31 (70)	.757	[a]	.014	-.03 (-.14, .07)	-0.55 (69)	.587	[a]	.004
CBT vs. WBT	-.09 (-.20, .01)	-1.52 (70)	.133	[a]	.046	-.03 (-.14, .09)	-0.37 (69)	.715	[a]	.006
Crystallized intelligence										
PBT vs. CBT	.03 (-.08, .13)	0.44 (70)	.659	[a]	.003	-.07 (-.20, .05)	-0.95 (69)	.346	[a]	.044
PBT vs. WBT	.08 (-.06, .21)	0.95 (70)	.344	[a]	.069	-.06 (-.17, .04)	-1.00 (69)	.323	[a]	.017
CBT vs. WBT	.08 (-.03, .18)	0.78 (70)	.440	[a]	.029	-.06 (-.18, .05)	0.10 (69)	.919	[a]	.030

Note. Effect sizes are Cohen's *d* for the trail-making test and the two self-report scales or differences in correlations for intelligence; t_{nil} / p_{nil} = Nil hypothesis test; t_{equ} / p_{equ} = Equivalence test; PBT = Paper-based assessment; CBT = Computerized assessment; WBT = Web-based assessment; Cryst. = Crystallized intelligence. The test of no difference in correlations is based on the difference in *p*-values of two normally distributed test statistics (see Counsell & Cribbie, 2015)

Figure 1

Mean Scores of the Trail-Making Test by Assessment Mode



Note. PBT = Paper-based assessment; CBT = Computerized assessment; WBT = Web-based assessment. Error bars indicate 95% confidence intervals.

Open Science

I report how I determined my sample sizes, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If I use inferential tests, I report exact p values, effect sizes, and 95% confidence or credible intervals.

Open Data: The information needed to reproduce all of the reported results are available at https://osf.io/qh4z9/. I confirm that there is sufficient information for an independent researcher to reproduce all of the reported results.

Open Materials: I confirm that there is sufficient information for an independent researcher to reproduce the reported methodology.

Preregistration of Studies and Analysis Plans: This study was not preregistered.