Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Clin Exp Neuropsychol. Author manuscript; available in PMC 2022 December 15.

Published in final edited form as:

J Clin Exp Neuropsychol. 2021 October ; 43(8): 786–795. doi:10.1080/13803395.2021.2002269.

The TestMyBrain Digital Neuropsychology Toolkit: Development and Psychometric Characteristics

Shifali Singh[a,b], **Roger W. Strong**[a,b], **Laneé Jung**[a,b], **Frances Haofei Li**[a,b], **Liz Grinspoon**[a,b], **Luke S. Scheuer**[a,b], **Eliza J. Passell**[a,b], **Paolo Martini**[a,b], **Naomi Chaytor**[c], **Jason R. Soble**[d,e], **Laura Germine**[a,b]

[a]Institute for Technology in Psychiatry, McLean Hospital, Belmont, MA, USA

[b]Department of Psychiatry, Harvard Medical School, Boston, MA, USA

[c]Elson S. Floyd College of Medicine, Washington State University, Spokane, WA, USA

[d]Department of Psychiatry, University of Illinois College of Medicine, Chicago, IL, USA

[e]Department of Neurology, University of Illinois College of Medicine, Chicago, IL, USA

Abstract

Introduction: To allow continued administration of neuropsychological evaluations remotely during the pandemic, tests from the not-for-profit platform, TestMyBrain.org (TMB), were used to develop the TMB Digital Neuropsychology Toolkit (DNT). This study details the psychometric characteristics of the DNT, as well as the infrastructure and development of the DNT.

Method: The DNT was primarily distributed for clinical use, with (72.8%) of individuals requesting access for clinical purposes. To assess reliability and validity of the DNT, anonymous data from DNT test administrations were analyzed and compared to a large, non-clinical normative sample from TMB.

Results: DNT test scores showed acceptable to very good split-half reliability (.68–.99). Factor analysis revealed three latent factors, corresponding to processing speed, working memory, and a broader general cognitive ability factor that included perceptual reasoning and episodic memory. Average test scores were slightly poorer for the DNT sample than for the TMB comparison sample, as expected given the clinical use of the DNT.

Conclusions: Initial estimates of reliability and validity of DNT tests support their use as digital measures of neuropsychological functioning. Tests within cognitive domains correlated highly with each other and demonstrated good reliability and validity. Future work will seek to validate DNT tests in specific clinical populations and determine best practices for using DNT outcome measures to assess engagement and psychological symptomatology.

CONTACT Shifali Singh ssingh@mclean.harvard.edu McLean Hospital, 115 Mill Street, Oaks Rm. 358A, Belmont, MA 02478.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Keywords

Digital neuropsychology; teleneuropsychology; assessment; psychometrics; cognition; remote testing

Introduction

The COVID-19 pandemic required clinicians to modify traditional neuropsychological assessment administration paradigms so they could continue conducting clinical evaluations while adhering to social distancing guidelines. Modifications to evaluations include wearing masks, using barriers to administer in-person assessments, offering remote assessments using videoconferencing (Marra et al., 2020), and using digital technologies to administer tests remotely (Singh & Germine, 2020).

Although modifications to test administration have been necessary for conducting neuropsychological evaluations during the pandemic, traditional tests were not originally normed using these modifications (e.g., wearing masks while administering verbal stimuli) and therefore are not currently being administered in accordance with standard practices. The results of two surveys completed in the beginning of the COVID-19 pandemic (Marra et al., 2020) suggest that neuropsychologists were divided on how to proceed given the COVID-19 restrictions. Those in private practice were significantly less likely to provide services than those in academic settings. Additionally, most clinicians (57%) continued to conduct outpatient evaluations, 31.3% of clinicians discontinued seeing any inpatients or outpatients, and 16.1% of all clinician respondents reported using neither a phone nor video, with no plan to use these services in the future.

As a result of the standardization issues, uncertain practice guidelines and personal and patient safety concerns, we began to receive numerous requests from clinicians, researchers, and educators familiar with our web-based, not-for-profit platform, TestMyBrain.org (TMB). They requested access to remote assessments through the platform, as traditional neuropsychological tests are not designed for remote administration. TMB uses a citizen-science approach, allowing remote data collection for individuals interested in taking self-administered cognitive tests, which are available for free on the front page. The large TMB normative database, which includes data from approximately 2.5 million individuals worldwide over the last 12 years, was appealing to those concerned about standardization and availability of appropriate normative data. In response to the high demand for remote neuropsychological assessments during the pandemic, we built a basic infrastructure for administering a select battery of empirically validated TMB cognitive assessments comparable to traditional neuropsychological instruments (Chaytor et al., 2021; Fortenbaugh et al., 2015; Germine et al., 2012, 2019; Hartshorne & Germine, 2015; Rutter et al., 2020; Vogel et al., 2020; Wilmer et al., 2012). We refer to this platform as the TMB Digital Neuropsychology Toolkit (DNT).

Given the continued interest in and use of the DNT, the first aim of this study was to report the psychometric properties of DNT tests based on DNT samples and the larger TMB normative database. The second aim was to discuss in detail the infrastructure and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

development of the DNT. This will be useful for readers who are currently using the DNT for clinical work, research, teaching, and/or have an interest in remote cognitive test development.

Method

TestMyBrain Digital Neuropsychology Toolkit infrastructure and development

The DNT was designed to be separate from the TMB research platform and provide greater clinical utility. The DNT infrastructure is unique in that it allows clinicians to design specific batteries that can be digitally sent to patients using a customized link. Additionally, through the DNT customized link, clinicians can retrieve raw data and standard scores normed on TMB participant data (Table 1). The TMB DNT, supported by McLean Hospital (part of Harvard Medical School) and the Many Brains Project (nonprofit 501c3), was publicly released on 2 April 2020. By 30 April 2020, there were 1,090 requests for access to the toolkit. All individuals requesting access were provided with instructions (Appendix) detailing how to use the DNT. In the last year, the number of requests and continued access has remained relatively stable (approximately 500 unique batteries per month) despite many clinicians returning to in-person assessment.

Although the DNT provides a means for administering remote assessment, there are several limitations. The toolkit relies on normative data from specific TMB tests and is a convenience-based normative sample (N = 135,265) with individuals who self-selected to complete online testing; this may offer a key difference with other normative samples and those randomly sampled from the population, as those who elected to complete online TMB tests may be more comfortable with online testing. Although specific tests have been used in clinical context, the DNT as a whole is not currently validated on known clinical samples, so clinicians are cautioned when using results for diagnostic purposes. Accessibility is another limitation, in that those who are less familiar with digital technologies such as smartphones or tablets may not be appropriate populations for the DNT. Moreover, although we did our best to select a wide range of tests, including commonly used mood and anxiety tests not included in analyses in this paper, the tests are by no means all-encompassing and are not able to fully assess cognitive domains/abilities as comprehensively as a traditional neuropsychological evaluation. For example, we do not include language tests, memory tests requiring free recall, or more broad measures of executive functioning to comprehensively capture heterogeneity of these cognitive domains. Online tests in general can also be limited in that many of them, like the DNT, do not allow for auditory administration or the recording of verbal responses. This limits the validity of data from patients with motor/visual limitations but might be beneficial for those with hearing loss or oral-motor issues. These types of measures may be integrated in future releases of the DNT. Finally, as with many digital platforms, traditional behavioral observations are limited, making it difficult to determine whether participants and/or patients are engaged in tasks they are completing remotely (Feenstra et al., 2017). Future work might address this topic, particularly given the unique ability of digital tests to provide item-level data that we can leverage to understand varying levels of engagement.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

For the TMB DNT, 11 tests were selected based on clinician demand and/or evidence of validity for the TMB research versions of the tests when compared with more traditional assessments (Chaytor et al., 2021; Fortenbaugh et al., 2015; Germine et al., 2012, 2019; Hartshorne & Germine, 2015; Rutter et al., 2020; Vogel et al., 2020; Wilmer et al., 2012). All tests show age-related differences in the expected direction and magnitude as traditional normative samples (Hartshorne & Germine, 2015). To enable the most flexibility for each clinician, we provided the clinician the capability to create custom batteries with tests selected for each individual patient.

We built the DNT to create a relatively easy way of building testing sessions for clinicians and retrieving data from testing sessions, while eliminating the need to store protected health information in our system in accordance with US HIPAA regulations. Our databases do not store or maintain any personally identifiable information (PII) data and, to date, no demographic data. All communications between the patient's client machine and the TMB server use transport layer security (TLS) encryption (designed to provide communications security over a computer network) and follow industry standards for secure communication. At no point in the process does our team receive any information that could connect any identifier with any particular individual. In addition, our toolkit provides no free text fields where anyone can provide PII data.

TMB participants

All normative data used to score DNT tests are from anonymous data collected on TestMyBrain.org (Table 1). For a comparison of the TMB and DNT normative data, along with our quality control rules, please see the Supplement. Self-reported demographic information includes age (12–99; $M = 28.8$, SD = 15.3), gender (female = 49.4%), and education (≤ 12 years = 29.4%). Participants were typically directed to the TMB page through Google searches and links posted by previous participants on social media. All participants included in the TMB normative dataset were recruited and consented through protocols approved by either the Harvard University Committee on the Use of Human Subjects or the Mass General Brigham (formerly Partners Healthcare) Institutional Review Board.

DNT requests

Most users requested access for clinical purposes (72.8%), followed by research (12.1%).

Please see Table 2 for additional descriptive statistics on the DNT.

Measures

Below are the tests included in the DNT, along with descriptions and score types used for psychometric analyses. All tests described below begin with detailed instructions and unscored practice trials to help ensure participants understand tasks before they begin. Figure 1 illustrates the tasks organized by cognitive domain. Please see the Supplement for histograms showing each measures' distributional properties.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Working memory/attention

TMB Digit Span (Forward and Backward; **Chaytor et al., 2021; Germine et al., 2012; Hartshorne & Germine, 2015).—**The patient is asked to recall sequences of visually presented digits of increasing length, in either the same order (Forward Digit Span; FDS) or the opposite order (Backward Digit Span; BDS). Digits are presented individually for 1000 milliseconds each. After the final digit is presented, participants are presented with text reading "Now press the numbers," at which point they are required to enter the span they had just viewed by typing the number series. Each span begins with only two numbers and increased to a maximum span of 11 numbers. Participants complete two trials at each span length; if the number span for at least one of those two trials is entered successfully, the span length is increased by one number. If two trials with the same span length are consecutively incorrect, the task is discontinued. Scores are recorded as the longest set length where at least one trial was completed successfully (FDS.Score, BDS.Score). The two subtests can be administered individually or together. Average completion time: 3.5 minutes each.

TMB Gradual Onset Continuous Performance Test (gradCPT; **Fortenbaugh et al., 2015).—**The patient presses a key when a city image appears and does not press when a mountain image appears. Images rapidly transition from one to the next, with mountains appearing only 10–20% of the time. The primary test scores (ones that are reported to clinicians) of interest and used for psychometric analyses are sensitivity (CPT.D; discrimination ability/d', a measure of target discriminability not impacted by response bias, where higher scores indicate better performance) and response bias (CPT.C; criterion, a measure of response bias where a larger value indicates greater response impulsivity, or tendency to press a key regardless of the picture type). Average completion time: 6 minutes.

Processing speed

TMB Simple Reaction Time (Simple RT; **Rutter et al., 2020).—**The patient presses a key (or taps the screen) whenever a green square appears. Inter-target intervals are generated based on an exponential distribution (to mitigate potential foreperiod effects). The primary test score of interest and used for psychometric analyses is median reaction time (SRT). Average completion time: 1 minute.

TMB Choice Reaction Time (Choice RT; **Rutter et al., 2020).—**The patient taps the button indicating the direction of the arrow that is a different color from the rest. The primary test scores of interest and used for psychometric analyses are median reaction time for correctly answered trials (CRT.RT) and percentage of correctly answered trials (CRT.ACC). Average completion time: 2.5 minutes.

TMB Digit Symbol Matching **(Chaytor et al., 2021; D'Ardenne et al., 2020; Hartshorne & Germine, 2015).—**Participants are presented with nine symbols, each of which are paired with a single digit between 1 and 3 (i.e., three symbols were paired with each digit); these pairings remain visible throughout the duration of the test. Individual probe symbols are sequentially presented above these pairings, to which participants respond by selecting the corresponding digit as quickly as possible; each probe symbol remains

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

visible until participants make a response. Scores are recorded as the total number of correct responses in 90 seconds. The primary test score of interest and used for psychometric analyses is number of correctly completed matches (DSM.score). Average completion time: 2 minutes.

***TMB Trail-Making Test Part A (*Trevino et al., Trevino, et al., in press).—**The patient connects a series of numbers in ascending order as quickly as possible using the mouse or touchscreen to drag lines from one number to the next. If an incorrect "trail" is made, the incorrect response turns red, and a red text warning notifies them that they are wrong and should go back to their last response before being allowed to continue. The primary test scores of interest and used for psychometric analyses are total time to complete the trail (TA.total), and median reaction time for each correctly sequenced number (TA.mRT). Average completion time: 2 minutes.

Memory

***TMB Verbal Paired Associates (*Hartshorne & Germine, 2015; Wilmer et al., 2012, 2010).—**Patients are visually presented with 25 unrelated word pairs one at a time and told that they will be asked to remember them later. Each word pair is presented individually for 6 seconds, with a 2 second interstimulus interval between pairs. After a delay of approximately 2 minutes (where the TMB Digit Symbol Matching test is completed), patients are sequentially presented with one word from each of the studied pairs and asked to recognize which word was previously paired with it by selecting the correct word from a list of four response options. The primary test score of interest and used for psychometric analyses is the number of words correctly selected (Verb). Average completion time: 5 minutes not including delay between encoding and recall trials, when other tasks can be completed.

***TMB Visual Paired Associates (*Passell et al., 2019).—**Participants are visually presented with 24 pairs of same-category, recognizable images (e.g., two distinct pictures of barns) and are told they will later be tested on which images were paired together. Each pair of images is presented for 5 seconds, with a 1.5 second interstimulus interval between pairs. After a delay of approximately 2.5 minutes (during which the TMB Choice Reaction Test is completed), participants are sequentially presented with one image from each of the studied pairs and asked to recognize which image was previously paired with it by selecting the correct image from a set of five response options. The primary test score of interest and used for psychometric analyses are the number of images correctly identified (Vis). Average completion time: 5 minutes not including delay between encoding and recognition trials, when other tasks can be completed.

Executive functioning

***TMB Trail-Making Test Part B (*Trevino et al., in press).—**The patient connects a series of numbers and letters in alternating ascending order. If an incorrect "trail" is made, this error is recorded, and the patient is required to go back to the previous number/letter and draw a path to the correct number/letter before being allowed to continue. The primary test scores of interest and used for psychometric analyses are total time to complete the trail

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(TB.total) and median reaction time for each correctly sequenced target (TB.mRT). Average completion time: 2 minutes.

Perceptual reasoning

TMB Matrix Reasoning **(Chaytor et al., 2021; D'Ardenne et al., 2020).—**On each trial, patients view a matrix of images with one image missing. Patients determine how the images are related and select the image that best completes the pattern from 5 response options. Trials become increasingly complex and continue until 3 consecutive errors are made or all 26 trials have been administered. The primary score of interest (Matrix) and used for psychometric analyses is the number of correct trials. Average completion time: 8 minutes.

Results

Reliability

Table 3 provides the number of participants, means, standard deviations and Spearman-Brown corrected split-half reliabilities for each test administered in both the DNT and TMB normative samples. For the TMB Forward and Backward Digit Span tests, the test design did not permit calculation of a split-half reliability. As an estimate of the reliability of these two tests, we report reliabilities computed from an alternate form of the tests, where a different sample of participants simultaneously completed two interleaved, independently scored digit spans (i.e., two interleaved forward or two interleaved backward digit spans; Chaytor et al., 2021); we report the correlation between these two spans as an estimate of the tests' reliability. For all other test outcomes, Spearman-Brown corrected split-half reliability was calculated (even versus odd trials for TMB Simple Reaction Time, Choice Reaction Time, Visual Paired Associates, Verbal Paired Associates, Matrix Reasoning, and the Gradual Onset Continuous Performance Test; first half of targets versus second half of targets for TMB Trail-Making Test Parts A and B; responses made during the first versus second half of testing duration for TMB Digit Symbol Matching).

In our primary analyses, if a participant completed a test more than once, only the first completion was included. For the majority of tests, fewer than 30 participants completed the test multiple times, prohibiting the calculation of test-retest reliability. An exception to this were the TMB Choice Reaction Time and Digit Symbol Matching tests, which were included twice in a subset of participants' batteries – both as a standalone test and as the intervening test between the study and test phases of the TMB Visual Paired Associates (Choice Reaction Time) and Verbal Paired Associates (Digit Symbol Matching) tests. In total, 925 participants competed the TMB Choice Reaction Time test multiple times, while 633 participants completed the Digit Symbol Matching test multiple times. We report three different metrics of test-retest reliability for these two tests: 1) Pearson correlation of the first and second completion by each participant; 2) intraclass correlation coefficients reflecting the consistency of scores between testing completions (ICC(C,1); McGraw & Wong, 1996); and 3) intraclass correlation coefficients reflecting the absolute agreement of scores between testing completions (ICC (A,1); McGraw & Wong, 1996). For TMB Choice Reaction Time, the primary outcome, median correct reaction time, had good test-retest

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

reliability (Pearson r = .81 [.78–83], ICC(C,1) = .80 [.78–.83], ICC(A,1) = .79 [.75–.83]). Accuracy, the secondary outcome, had poor test-retest reliability (Pearson r = .48 [.43–53], ICC(C,1) = .46 [.41–.51], ICC(A,1) = .45 [.40–.50]), likely due to ceiling effects. For TMB Digit Symbol Matching, total score had good test-retest reliability (Pearson r = .88 [.86–.90], ICC (C,1) = .84 [.82–.86], ICC(A,1) = .73 [.17–.88]).

Exploratory factor analysis

We conducted an exploratory factor analysis to assess latent groupings between tests available in the DNT. For tests with multiple outcomes, we included only each test's primary outcomes measure: total Reaction Time for Trails A and B, Median Correct Reaction Time for Choice Reaction Time, and Sensitivity (D) for CPT. We first performed an exploratory factor analysis using only data from participants who completed all 11 cognitive tests (n = 538). A parallel analysis (Zwick & Velicer, 1986) using the fa. parallel() function of the *psych* package in R (Revelle, 2017) suggested three latent factors (see Supplement for scree plot); therefore, we conducted an exploratory factor analysis with three latent factors and direct oblimin rotation (to allow correlations between factors) using the *psych* package's fa() function. Factor loadings and structure are presented in Figure 2. Based on this factor structure, we determined three distinct cognitive domains: processing speed, working memory, and general cognitive ability. Processing speed included Trail Making Test Parts A and B, Choice Reaction Test, and Simple Reaction Time. Working memory included Forward and Backward Digit Span, and general cognitive ability included Verbal and Visual Paired Associates, the Gradual Onset Continuous Performance Test, Matrix Reasoning, and Digit Symbol Matching. Notably, Digit Symbol Matching loaded similarly on the general cognitive ability and processing speed factors, likely because it is a speeded reaction time task with a memory component (Joy et al., 2003). As a sensitivity analysis, we randomly divided participants who completed all measures into two groups (n = 269 each), on each of which we separately performed an exploratory factor analysis. The factor loadings were very similar for each half of participants (see Supplement), demonstrating that this factor structure is robust.

We additionally performed an exploratory factor analysis that included data from all participants, using full information maximum likelihood estimation to compute the correlation matrix among test outcomes (via the corFiml() function of the *psych* package). See Table 1 for the number of participants who completed each measure, and Figure 3 for the number of participants who completed each pair of measures. We again performed an exploratory factor analysis with three latent factors and direct oblimin rotation, which produced nearly the same factor structure as found using only the subset of participants who completed all 11 tests; the only test whose highest factor loading changed was Digit Symbol Matching, which loaded more heavily on the processing speed factor than the general cognitive ability factor (see Supplement).

Discussion

The TMB DNT is a clinician-friendly tool that can be used to facilitate remote neuropsychological evaluations. The large number of DNT test batteries continually created

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

indicates clinicians are finding this battery of digital measures useful even as many practices are returning to in-person assessments. This underscores the importance of studies detailing the development of digital tools with psychometric studies evaluating their validity and reliability. This study provides a basic initial report of the psychometric characteristics, development, and infrastructure of DNT tests. Although previous studies have provided psychometric data for all DNT tests in a research context (e.g., Chaytor et al., 2021; Germine et al., 2012), this study is the first to examine TMB tests specifically in the context of the DNT in a largely clinical sample. To protect patient confidentiality, there are no patient demographic or medical information collected, nor are there any data on the types of evaluations or clinical contexts for which the DNT was used; therefore, the results of this study are limited to test performance data collected from individuals who were provided links to DNT batteries, without screening for demographic variables or clinical history.

The DNT demonstrates acceptable to very good split-half reliability (Ursachi et al., 2015). Results from the factor analysis suggest that the measures included in the DNT load on 3 distinct factors, roughly corresponding to 1) processing speed, 2) working memory, and 3) general cognitive ability. This structure aligns with expectations, particularly in terms of speeded tasks loading on the same factor. Consistent with this, a recent study using TMB (Trevino et al., in press) identified a 5-factor structure, with Forward and Backward Digit Span as one factor, and Gradual Onset Continuous Performance Test as its own factor, also separated from speed-based tasks. Additionally, the cognitive domains in the widely-used Wechsler Adult Intelligence Scale (Wechsler, 2008), for example, include perceptual reasoning, processing speed, and working memory factors; these factors map on well to our general cognitive ability, processing speed, and working memory factors, respectively. The current study confirms the expectations that TMB DNT produced data similar in overall scores, reliability, and correlations to data from our TMB normative dataset, with test scores being positively correlated (Figure 3) within specific cognitive domains like processing speed (convergent validity) and having lower correlations with tests in other cognitive domains (divergent validity).

Given that self-administered, online tests have emerging evidence of comparable construct and ecological validity to traditional neuropsychological tests (e.g., Chaytor et al., 2021), it is likely that digital tools will become increasingly used in neuropsychological testing both during and beyond the pandemic. As discussed, there are several limitations associated with using the DNT and digital cognitive tests in general, including the limited range of available digital cognitive tests and normative data. With the DNT specifically, we do not currently have data on the clinical populations reflected in our sample, which limits clinical utility. Looking forward, it will be important to determine which demographic variables are most relevant for each DNT test and use that information to develop appropriate demographically adjusted normative data. Future work will also need to validate DNT tests on diverse clinical samples, identify appropriate intra-test measures of performance validity, and cross-validate those measures with well-validated criterion performance validity tests. In doing this, we can help better objectively identify invalid test performance levels within specific digital cognitive tests.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Despite the DNT's current limitations, there are several options for incorporating the toolkit effectively in clinical practice. Instead of using it as a standalone battery, clinicians may opt to choose tests that supplement traditional cognitive tests in a neuropsychological evaluation. Based on our findings, processing speed measures may be especially useful in a teleneurop sychology setting, where more precise and nuanced measurements of processing speed are needed. Ultimately, although remote cognitive testing is only recently being incorporated in traditional clinical neuropsychological evaluations and requires continued validation, it holds significant promise in adequately reaching clinicians and researchers whoseek to supplement traditional tests with digital measures of cognitive functioning.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

The author(s) reported there is no funding associated with the work featured in this article.

References

Chaytor NS, Barbosa-Leiker C, Germine LT, Fonseca LM, McPherson SM, & Tuttle KR (2021). Construct validity, ecological validity and acceptance of self-administered online neuropsychological assessment in adults. The Clinical Neuropsychologist, 35(1), 148–164. 10.1080/13854046.2020.1811893 [PubMed: 32883156]

D'Ardenne K, Savage CR, Small D, Vainik U, & Stoeckel LE (2020). Core neuropsychological measures for obesity and diabetes trials: Initial report. Frontiers in Psychology, 11, 1–13 . 10.3389/fpsyg.2020.554127 [PubMed: 32038435]

Feenstra HEM, Vermeulen IE, Murre JMJ, & Schagen SB (2017). Online cognition: Factors facilitating reliable online neuropsychological test results. The Clinical Neuropsychologist, 31(1), 59–84. 10.1080/13854046.2016.1190405 [PubMed: 27266677]

Fortenbaugh FC, DeGutis J, Germine L, Wilmer JB, grosso m., russo k., & esterman m. (2015). sustained attention across the life span in a sample of 10,000: Dissociating ability and strategy. Psychological Science, 26 (9), 1497–1510. 10.1177/0956797615594896 [PubMed: 26253551]

Germine L, Nakayama K, Duchaine BC, Chabris CF, Chatterjee G, & Wilmer JB (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. Psychonomic Bulletin & Review, 19(5), 847–857. 10.3758/s13423-012-0296-9 [PubMed: 22829343]

Germine L, Reinecke K, & Chaytor NS (2019). Digital neuropsychology: Challenges and opportunities at the intersection of science and software. The Clinical Neuropsychologist, 33(2), 271–286. 10.1080/13854046.2018.1535662 [PubMed: 30614374]

Hartshorne JK, & Germine LT (2015). When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. Psychological Science, 26(4), 433–443. 10.1177/0956797614567339 [PubMed: 25770099]

Joy S, Fein D, & Kaplan E (2003). Decoding digit symbol: Speed, memory, and visual scanning. Assessment, 10(1), 56–65. 10.1177/0095399702250335 [PubMed: 12675384]

Marra DE, Hoelzle JB, Davis JJ, & Schwartz ES (2020). Initial changes in neuropsychologists clinical practice during the COVID-19 pandemic: A survey study. The Clinical Neuropsychologist, 34(7–8), 1251–1266. 10.1080/13854046.2020.1800098 [PubMed: 32723158]

McGraw KO, & Wong SP (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1(1), 30–46. 10.1037/1082-989X.1.1.30

Osborne JW, & Waters E (2002). Four assumptions of multiple regression that researchers should always test. Practical assessment, research, and evaluation, 8(1), 2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Passell E, Dillon DG, Baker JT, Vogel SC, Scheuer LS, Mirin NL, Rutter LA, Pizzagalli DA, & Germine L (2019). Digital cognitive assessment: Results from the TestMyBrain NIMH Research Domain Criteria (RDoC) field test battery report 10.31234/osf.io/dcszr

Revelle W (2017). Psych: Procedures for Personality and Psychological Research Evanston, Illinois, USA: Northwestern University. https://CRAN.R-project.org/package=psych

Rutter LA, Vahia IV, Forester BP, Ressler KJ, & Germine L (2020). Heterogeneous indicators of cognitive performance and performance variability across the lifespan. Frontiers in Aging Neuroscience, 12, 62. 10.3389/fnagi.2020.00062 [PubMed: 32210793]

Singh S, & Germine L (2020). Technology meets tradition: A hybrid model for implementing digital tools in neuropsychology. International Review of Psychiatry, 33(4), 382–393. 10.1080/09540261.2020.1835839 [PubMed: 33236657]

Trevino M, Zhu X, Lu Y, Scheuer LS, Passell E, Huang GC, Germine LT, & Horowitz TS (in press). How do we measure attention? Using factor analysis to establish construct validity of neuropsychological tests. Cognitive Research: Principles and Implications, 6(1), 1–26. 10.1186/s41235-021-00313-1

Ursachi G, Horodnic IA, & Zait A (2015). How reliable are measurement scales? External factors with indirect influence on reliability estimators. Procedia Economics and Finance, 20, 679–686. 10.1016/S2212-5671(15)00123-9

Vogel SC, Esterman M, DeGutis J, Wilmer JB, Ressler KJ, & Germine LT (2020). Childhood adversity and dimensional variations in adult sustained attention. Frontiers in Psychology, 11, 691. 10.3389/fpsyg.2020.00691 [PubMed: 32362858]

Wechsler D (2008). WAIS-IV administration and scoring manual (Canadian) The Psychological Corporation.

Wilmer JB, Germine L, Chabris CF, Chatterjee G, Gerbasi M, & Nakayama K (2012). Capturing specific abilities as a window into human individuality: The example of face recognition. Cognitive Neuropsychology, 29(5–6), 360–392. 10.1080/02643294.2012.753433 [PubMed: 23428079]

Wilmer JB, Germine L, Chabris CF, Chatterjee G, Williams M, Loken E, Nakayama K, & Duchaine B (2010). Human face recognition ability is specific and highly heritable. Proceedings of the National Academy of Sciences, 107(11), 5238–5241. 10.1073/pnas.0913053107

Zwick WR, & Velicer WF (1986). Comparison of five rules for determining the number of components to retain. Psychological Bulletin, 99(3), 432–442. 10.1037/0033-2909.99.3.432

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

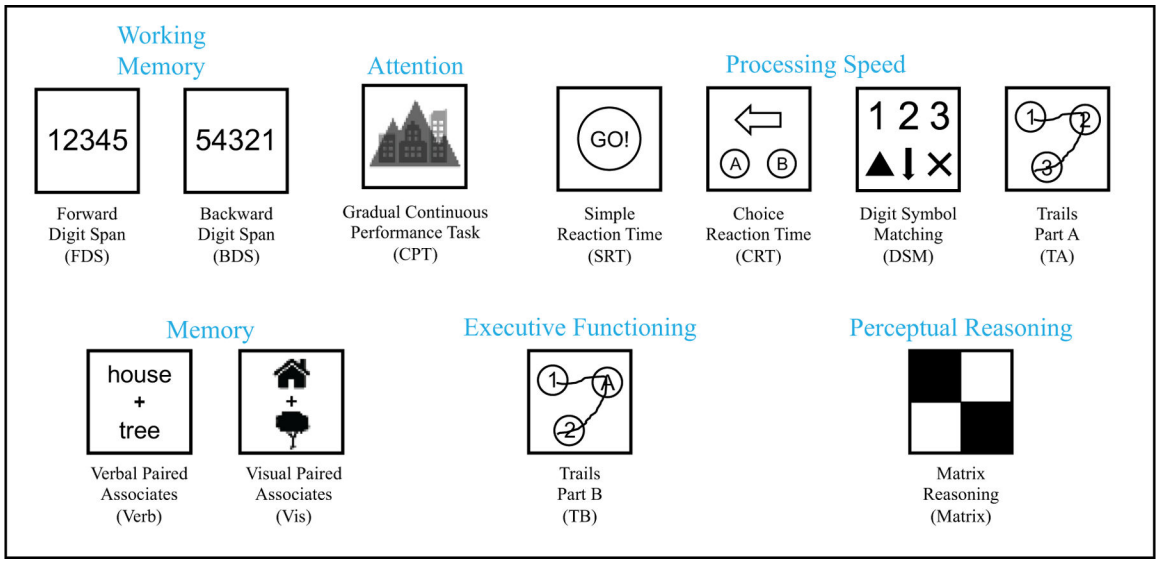

Figure 1.
Visualization of tests included in the DNT.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

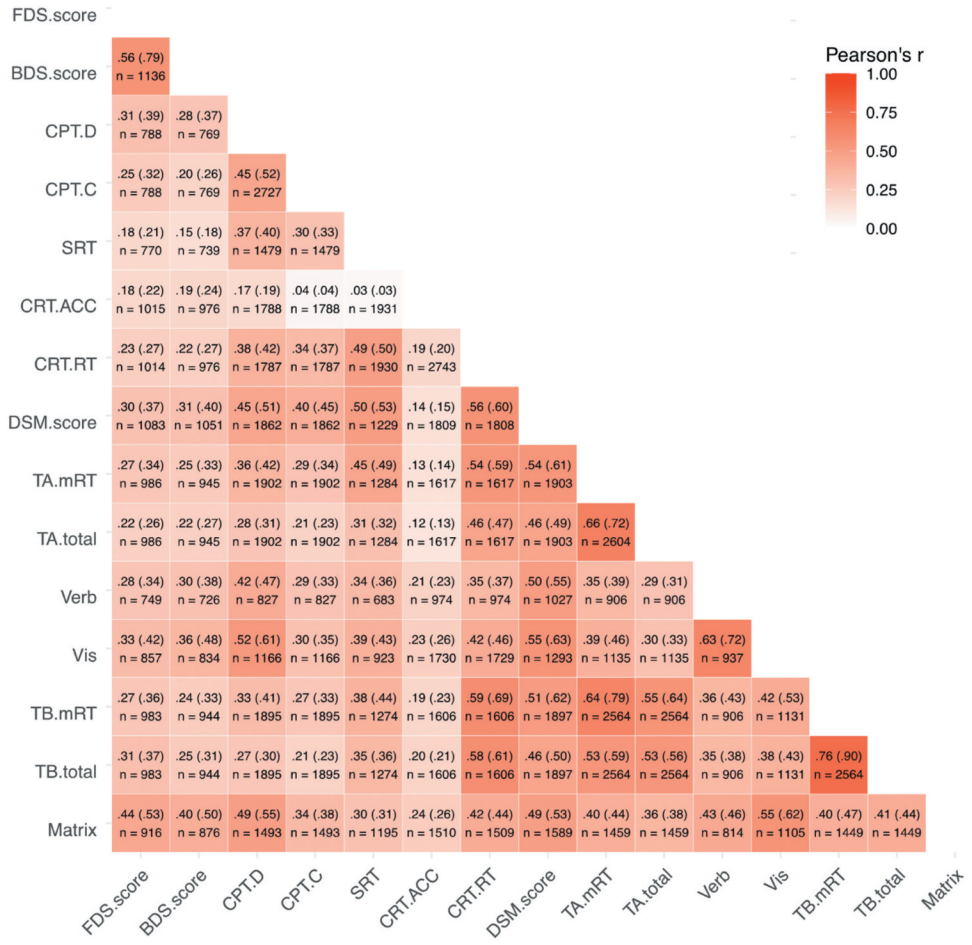

Figure 2.

Factor Loading Strengths for Cognitive Measures. Reaction time measures were reverse scored, so that higher scores were indicative of better performance for all measures. TB.total = Trail Making Test – Part B scores, based on total time to complete the trail (in milliseconds). TA.total = Trail Making Test – Part A scores, based on total time to complete the trail (in milliseconds). CRT.RT = Choice Reaction Time scores, median response time (correct trials) in milliseconds. SRT = Simple Reaction Time scores, or median response time in milliseconds. FDS. score = Forward Digit Span scores, or maximum sequence length accurately recalled. BDS.score = Backward digit span scores, or maximum sequence length accurately recalled. Vis = Visual paired associates memory scores: total number correct (24 maximum). DSM = Digit symbol matching scores, or number of digits and symbols correctly matched in 90 seconds. CPT.D = Gradual onset continuous performance test discrimination score, based on signal detection theory. Matrix = Matrix reasoning scores: total number correct (36 maximum). Grey bars show the loading strength of each measure on each factor. Cumulative proportion of variance explained by the three factors was .51. Total (non-unique) proportion of variance explained by each factor was .31 for Processing Speed, .19 for working memory, and .29 for General Cognitive Ability. Values at top of figure show correlations between factors.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

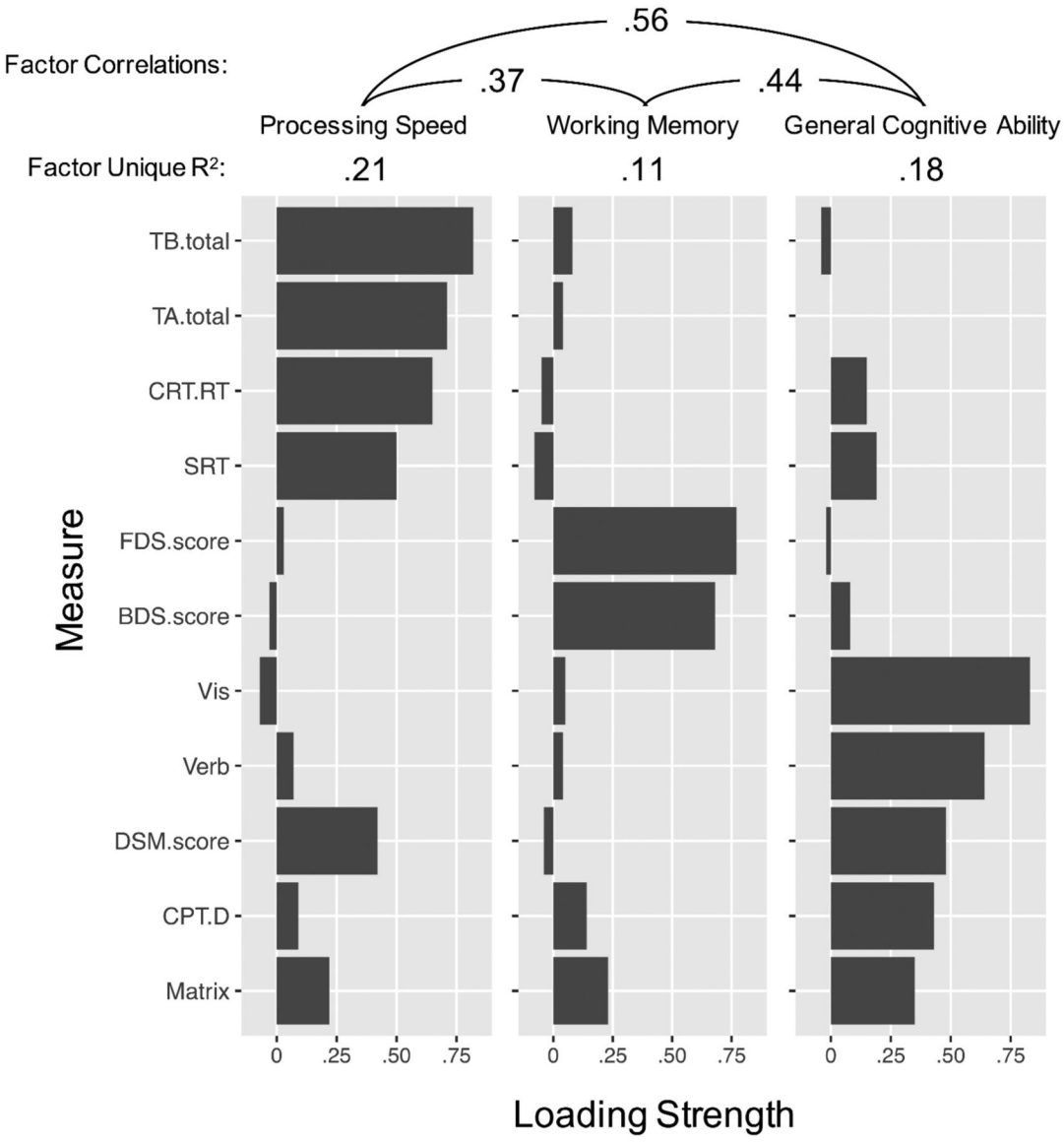

Figure 3.

Correlations between DNT test scores. Each cell shows the observed Pearson correlation between test scores, the adjusted correlation between scores accounting for reliability (in parentheses; Osborne, 2002), and the number of DNT participants for each pair of scores. SRT = Simple Reaction Time scores, or median response time in milliseconds. CRT.RT = Choice Reaction Time scores, median response time (correct trials) in milliseconds. CRT.ACC = Choice Reaction Time accuracy, or proportion correct responses. BDS. score = Backward digit span scores, or maximum sequence length accurately recalled. FDS.score = Forward Digit Span scores, or maximum sequence length accurately recalled. DSM = Digit symbol matching scores, or number of digits and symbols correctly matched in 90 seconds. Matrix = Matrix reasoning scores: total number correct (36 maximum). CPT.C = Gradual onset continuous performance test criterion score, based on signal detection theory. Higher scores indicate faster or more impulsive responses. CPT. D = Gradual onset

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

continuous performance test discrimination score, based on signal detection theory. Higher scores indicate more accurate responses (better performance). Vis = Visual paired associates memory scores: total number correct (24 maximum). Verb = Verbal paired associates memory scores: total number correct (25 maximum). TA.total = Trail Making Test – Part A scores, based on total time to complete the trail (in milliseconds). TA.mRT = Trail Making Test – Part A median response time for each correct response. TB.total = Trail Making Test – Part B scores, based on total time to complete the trail (in milliseconds). TB.mRT = Trail Making Test – Part B median response time for each correct response.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1.

TestMyBrain demographic data.

	Total (*n* = 135,265)	**Range**
Age, mean years (SD)	28.8 (15.3)	12–99
Age group, n (%)		
12–19	49,393 (36.5)	
20 – 29	38,321 (28.3)	
30–39	18,598 (13.7)	
40 – 49	11,717 (8.7)	
50–59	9175 (6.8)	
60 – 69	5539 (4.1)	
70–79	1982 (1.5)	
80 – 89	474 (0.4)	
90–99	66 (0.05)	
Female, n (%)	66,754 (49.4)	
Education, mean years (SD)	14.3 (2.9)	6–18
Education, n (%)		
<12 years	9437 (7)	
12 years	29,851 (22.1)	
12–15 years	27,478 (20.3)	
16 years	26,099 (19.3)	
18 years	24,263 (17.9)	
Did not disclose	18,137 (13.4)	
In the US, n (%)	51,657 (38)	
Outside the US, n (%)	81,496 (60)	
No location information, n (%)	2,112 (2)	
English as a native language, n (%)	92,647 (68)	
English not native language, n (%)	42,618 (32)	

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2.

Descriptive statistics on TestMyBrain digital neuropsychology toolkit*.

Assessment request data	*n*
Total requests	1,316
Number of unique batteries created	6,182
Pages created/week on average	119
Assessment batteries/day	17
Completed assessment batteries	4,714
User Data	%
Reporting clinical use	72.8
Reporting research use	12.1
Reporting training or educational use	10.3
Reporting personal use	1.2
Combination of above or unsure of use	3.7

* Data collected between 04/03/2020 and 03/05/2021.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3.

Descriptive characteristics of DNT and TMB samples.

	Digital Neuropsychology Toolkit Sample			TMB Normative Sample		
Test Score	**N**	**Mean (SD)**	**Split-Half Reliability**	**N**	**Mean (SD)**	**Split-Half Reliability**
Forward Digit Span – Score	1213	6.31 (1.82)	*a*	9858	6.43 (1.58)	*a*
Backward Digit Span – Score	1149	5.22 (2.04)	*a*	6269	5.33 (1.58)	*a*
GradCPT – D prime	2727	2.46 (0.92)	.86	20,197	2.88 (0.85)	0.86
GradCPT – Criterion	2727	0.61 (0.50)	.86	20,197	0.82 (0.37)	0.82
Simple RT – RT (ms)	1982	367.47 (146.0)	.99	49,668	303.21 (63.06)	0.97
Choice RT – Accuracy[b]	2745	.94 (.13)	.93	17,121	.95 (.08)	0.73
Choice RT – Correct RT (ms)[b]	2743	1172.86 (688.05)	.97	17,121	895.6 (257.47)	0.96
Digit Symbol Matching – Score	2642	44.02 (13.33)	.90	36,000	51.87 (12.20)	0.91
Trail-Making Test Part A – Median RT (ms)	2604	869.20 (405.01)	.87	21,278	691.09 (230.84)	0.86
Trail-Making Test Part A – Total Time (ms)	2604	35,551(30,488)	.97	21,278	27,420 (11,038)	0.94
Verbal Paired Associates – Score	1027	16.90 (6.65)	.92	8496	18.21 (5.63)	0.89
Visual Paired Associates – Score	1730	14.44 (5.27)	.84	9280	15.97 (4.53)	0.78
Trail-Making Test Part B – Median RT (ms)	2564	1407.88 (1235.40)	.76	21,343	1065.43 (410.12)	0.73
Trail-Making Test Part B – Total Time (ms)	2564	58,163 (65,371)	.94	21,343	41,776 (16,774)	0.93
Matrix Reasoning – Score	2058	24.22 (7.59)	.94	3674	26.43 (5.87)	0.91

[a]The design of the Forward and Backward Digit Span tests did not allow for computation of split-half reliability. In an alternate version of these tests where separate participants completed two interleaved, independently scored digit spans, reliability was .73 for the Forward Digit Span Test, and .68 for the Backward Digit Span Test (Chaytor et al., 2021).

[b]Differing sample sizes because two individuals did not produce any correct trials, so median correct reaction time calculation would be infeasible.