Cognitive Disabilities, Technological Flexibility, and The NIH Toolbox – Cognitive Battery: A Literature Review

Angie Donna Nikoleychuk

Department of Psychology, University of Lethbridge

Author Note

For the fulfillment of NEURO 3990 - Functional Design, Cognition, and Decision-Making

Abstract

The National Institute of Health Toolbox – Cognitive Battery was designed to provide researchers and professionals with an effective, accurate, low-cost method of assessing various aspects of cognitive function and cognitive processes. While its validity is generally acceptable and above, it is not accurate enough in certain circumstances to be used as a sole measurement. Some of these inaccuracies and lack of validity and reliability may also be due in part to shortfalls in the devices or design. This literature review examines the body of literature that currently exists and identifies areas that could benefit from further study.

Keywords: NIH Toolbox, cognitive assessment, executive function, working memory, language, attention, cognitive function

Cognitive Disabilities, Technological Flexibility, and The NIH Toolbox – Cognitive Battery: A Literature Review

The National Institute of Health Toolbox – Cognitive Battery (NIHTB-CB) was developed to assess cognitive abilities using iPad technology, particularly for large longitudinal and epidemiologic studies (Gershon et al., 2010). Specifically, the NIHTB-CB assesses executive function and attention, episodic memory, language, working memory, and processing speed across the lifespan to provide insights into how cognitive function changes through development and with age (Weintraub et al., 2013). While the tool has adequate to excellent validity for basic assessments across a general population in age groups ranging from 3 to 85 (Akshoomoff et al., 2013; Bauer, Dikmen, et al., 2013; Bauer, Zelazo, et al., 2013; Carlozzi et al., 2013; Gershon et al., 2013; D. M. Mungas et al., 2013; Tulsky et al., 2013; Weintraub et al., 2013; Zelazo et al., 2013), it was not developed specifically for populations diagnosed with cognitive disorders such as attention deficit hyperactivity disorder (ADHD) or for use with alternative technology. There has also been very little research into how the visual design of the tool and the hardware may affect results in various conditions. This literature review provides an overview of the original development of the NIHTB-CB and studies testing the toolbox under various conditions to understand which conditions may need to be controlled, under what circumstances may the tests need to be adapted, and which areas would benefit from further study.

Overview of the NIHTB-CB

The NIHTB-CB is a set of cognitive assessments available for the iPad to measure eight areas, which were developed from an existing set of traditional measurement tools or prior

established research or practices. These are outlined in Table 1. Generally, the NIHTB-CB offers a full range of tests, all with a generally moderate to high validity.

In addition to the domain-specific tests, the NIHTB-CB also offers Composite Crystalized, Fluid, and Overall Cognition Scores. The two-component theory of intellectual development is centered on the idea that intellectual development is organized into fluid and crystallized abilities, which change through the lifespan (Cattell, 1971; Horn 1968, 1970; Li et al., 2004). Fluid abilities, which are processes utilized when engaged in problem-solving, responding, and adapting to novel situations, are based heavily on biological processes as opposed to learning experiences. The development of fluid abilities increases through childhood to support the knowledge acquisition necessary for crystallized abilities. They peak in early adulthood, before declining with age and are highly susceptible to age-related degradation. Crystallized abilities, which include knowledge and skills based on life experiences and influenced by culture and education, improve throughout life until middle adulthood when they stabilize. They are more resilient to age and health-related degradation (Akshoomoff et al., 2013).

Initial validity found the NIHTB-CB to be reliable for children aged 3–15. However, researchers did note the small sample size. Test-retest reliability matched that of other gold-standard tests. Longitudinal data found a modest practice effect due to the nature of the tasks, so it must be controlled for when using this testing method. Regardless, the NIHTB-CB was shown to be an effective tool for tracking developmental and aging effects in ages 3 to 85. However, this conclusion isn't the consensus. When compared against comprehensive neuropsychological assessments for adults ranging in age from 20 to 66, Ott et al., 2022 found the NIHTB-CB had a

construct validity of merely poor to adequate for all six cognitive domains, as well as the fluid composite sore.

In addition, researchers behind the NIHTB-CB found no gender or SES effects. However, ethnic differences were discovered in some tests, so, therefore, this aspect could be a confounding factor (Akshoomoff et al., 2013; Kairys et al., 2021). Lastly, NIHTB-CB Composite scores were strongly related to maternal ratings for children's school performance (D. M. Mungas et al., 2013). Some studies have reported low reliability for Global Cognition, and Composite scores had a poor model fit and unsatisfactory reliability for ethnically diverse children, even when the scores were divided into three dimensions. (D. M. Mungas et al., 2013; Neumann et al., 2021).

The NIHTB-CB and Cognitive Disorders

Every cognitive disorder, disability, or disease is accompanied by unique challenges and deficiencies. And very few of these issues have been measured and tested for accuracy and validity with the NIHTB-CB. However, by examining a select few, it is possible to identify possible challenges, sensitivities, and complications that may be encountered by individuals, particularly if they are undiagnosed.

Alzheimer's Disease and Age-Related Cognitive Decline

While still a small body of work, several studies using the NIHTB-CB in individuals with Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI) have found no significant issues; the assessment tool was generally well accepted and valid (Bhaumik et al., 2020; Marques-Costa et al., 2020). However, many note that further research is necessary. Ma et al., 2020, found that the fluid cognition composite was a sensitive measure when detecting AD-

related cognitive impairment, while the crystallized cognition composite could be an efficient measure for cognitive reserve (Ma et al., 2021). If fluid cognitive ability scored below the crystallized cognitive ability, or the two scores varied significantly, it could potentially act as an early warning sign of AD and be associated with an increase in beta-amyloid deposits and cortical thickness (Ma et al., 2021; McDonough et al., 2016).

Special consideration needs to be taken when assessing AD and older individuals. The Flanker, DCCS, and Processing Speed tests include reaction time in their scores. Since older individuals may have motor impairments or become fatigued more easily, these confounding factors could account for some of the age-related effects often observed (Ma et al., 2021). Ma et al., 2021 also noted that "the tests are too challenging for individuals with AD and insensitive at the lower end of memory function, suggesting potentially limited utility for this population." It is interesting to note that, when iPads and computer-based cognitive programs are used to improve cognitive function, fatigue and difficulty were not mentioned as a limitation. However, in many of these studies, the task difficulty is often adjusted after determining a baseline. They are also "gamified" (Shao et al., 2015). While fun isn't the goal of the NIHTB-CB, it may be possible to integrate the assessments into a game-like context to improve completion rates.

Autism Spectrum Disorder and Intellectual Disability

Autism Spectrum Disorder research is overshadowed by the lack of valid, accessible assessment tools, which makes the NIHTB-CB attractive. While studies have found the various tests in the Toolbox to have moderate to strong reliability and convergent validity, there are some issues that would have to be addressed for the tool to become a reliable measurement tool (Jones et al., 2022; Shields et al., 2020; Solomon et al., 2020). Solomon et al., 2020 noted that ASD participants scored lower on fluid cognitive tasks, but the same as neuro-typical participants for

crystallized cognition. She found that ASD participants either scored low across all tasks or scored lower in fluid cognition, while all seemed to score lower on adaptive functioning tests.

Jones et al., 2022 observed similar results, but noted that older ASD children performed worse on the Flanker and Toolbox Picture Sequence Memory tests, suggesting the images used may not be adequate to keep them engaged. Lower scores in the Toolbox Picture Sequence Memory tests in autistic adults suggests that the way the test presents images could be problematic. During the test, images are presented sequentially, but they were relational, with points given for each pair placed in the correct order. However, autistic adolescents struggle to recognize images when presented without context or other images (Jones et al., 2022). The most common issue noted in many of the studies, however, was abandonment and low completion rates. As noted in Shields et al., 2020 and Jones et al., 2022, adaptations and adjustments will need to be made to assess ASD individuals with a mental age of approximately 5 years of age.

Other Cognitive Disorders

Individuals with mild cognitive impairment and above-average intellectual abilities pose a unique challenge for the NIHTB-CB since they often fail to obtain low test scores (Iverson & Karr, 2021). In this instance, adjusting for education, base rate analysis, and identifying higher than average scores may be the key to accurate assessment (Holdnack et al., 2017; Iverson & Karr, 2021). While examining working memory deficits in children with ADHD, Jusko et al. found the List Sorting Working Memory task was unable to identify participants with ADHD. These results may be due to the fact that the task relies on phonological working memory, which these individuals often rely on (Jusko et al., 2021). However, the List Sorting Working Memory task, along with the Flanker Inhibitory Control and Attention Test and the Dimensional Change

Card Sort Test had results similar to their traditional counterparts and were associated with clinical and self-reported anxiety and depression in children (Kavanaugh et al., 2020).

Cognitive Assessment, Technology, and Design

Mobile devices have the potential to revolutionize cognitive research and assessment, but it isn't without their pitfalls and intricacies (Dufau et al., 2011). Design can influence choices and make users more or less cooperative (Jiang & Fang, 2020). A user interface or website design can also significantly affect the accuracy, acceptance, and completion rates (Fox et al., 2020; Jones et al., 2022). Causing a user to wait, providing unclear labels, and failing to design software that allows users to understand intuitively how to complete a task can lead to high disengagement and abandonment rates (Nielsen, 1999). In short, optimal designs should be intuitive and not require additional explanation, minimize the amount of extraneous detail the user needs to remember, have the least number of steps possible, contain friction to avoid slips and mistakes, and contain informative feedback loops (Norman, 2013).

Habit and familiarity may also affect the outcomes determined by software-based assessment tools. Does a mobile device act as a contextual cue, causing participants to anticipate notifications, rings, or other signals they've become accustomed to? While mobile devices have varied results on cognition and human behaviour, one study found that participants who simply didn't have their phones on them scored higher on subsequent anxiety tests (Wilmer et al., 2017). Smartphone use has also been shown to decrease inhibition and decision-making ability (Warsaw et al., 2021). Does familiarity with the operation of a device play on the practice effect, decreasing response and navigation time? There currently are no answers to these and many other questions. Finding answers to these questions should be the next step.

Because familiarity and the technology can have profound negative and positive effects on reaction time, accuracy, usability, and other aspects, further research should present the technology in an unfamiliar format. Validity should also be tested across devices and setups including the iPad, smartphone, and desktop or laptop computer. Lastly, we know that seemingly insignificant details such as drop shadows to indicate a clickable button can relay an abundance of information to the user without their conscious knowledge (Norman, 2013). These details should be systematically identified and tested to ensure the NIHTB-CB is as accurate and sensitive as possible. Once these research questions have been answered, gamification should be explored, particularly for the assessment of younger groups.

Conclusion

While many studies have found the NIHTB-CB has excellent validity and reliability, it isn't unanimous and could use further improvements to increase the suitability and accuracy of the assessment tool. While many researchers have mentioned less than ideal validity or reliability, the inconsistencies haven't been extensively explored in many instances. To date, no studies have examined the influence the devices or the software design might have on the results, completion rates, or usability. Therefore, it is possible that the design of the software and hardware could play a role. Further testing is also needed to explore gamification and the benefits of newer concepts such as co-design principles (Fox et al., 2020).

Reference List

Akshoomoff, N., Beaumont, J. L., Beaumont, J. L., Patricia J. Bauer, Bauer, P. J., Dikmen, S. S., Richard Gershon, Gershon, R., Gershon, R., Mungas, D. M., Jerry Slotkin, Jerry Slotkin, Slotkin, J., Tulsky, D. S., Weintraub, S., Zelazo, P. D., Philip David Zelazo, Zelazo, P. D., & Heaton, R. K. (2013). VIII. NIH Toolbox Cognition Battery (CB): Composite scores of crystallized, fluid, and overall cognition. *Monographs of The Society for Research in Child Development*, *78*(4), 119–132. https://doi.org/10.1111/mono.12038

Alvseike, H., & Brønnick, K. (2012). Feasibility of the iPad as a hub for smart house technology in the elderly: Effects of cognition, self-efficacy, and technology experience. *Journal of Multidisciplinary Healthcare*, *5*, 299–306. https://doi.org/10.2147/jmdh.s35344

Amber N Schaefer, & Nicholls, C. J. (2021). A-30 Measures of attention and inhibitory control: Comparing the TOVA to the NIH Toolbox Flanker Test in children and adolescents with ADHD. *Archives of Clinical Neuropsychology*, *36*(6), 1071–1071. https://doi.org/10.1093/arclin/acab062.48

Anderson, V., Anderson, V., Beauchamp, M. H., & Beauchamp, M. (2017). *PEERS, Developing an iPad-delivered social assessment tool for children and adolescents, NHMRC*. *2017*(10), 80–82. https://doi.org/10.21820/23987073.2017.10.80

Anmarkrud, Ø., Andresen, A., & Bråten, I. (2019). Cognitive load and working memory in multimedia learning: Conceptual and measurement issues. *Educational Psychologist*, *54*(2), 61–83. https://doi.org/10.1080/00461520.2018.1554484

Bauer, P. J. (2005). New developments in the study of infant memory. In *Handbook of Research Methods in Developmental Science, D.M. Teti (Ed.)*. https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470756676.ch23

Bauer, P. J. (2006). Constructing a past in infancy: A neuro-developmental account. *Trends in Cognitive Sciences*, *10*(4), 175–181. https://doi.org/10.1016/j.tics.2006.02.009

Bauer, P. J. (2007). Recall in infancy: A neurodevelopmental account. *Current Directions in Psychological Science*, *16*(3), 142–146. https://doi.org/10.1111/j.1467-8721.2007.00492.x

Bauer, P. J. (2014). *Remembering the Times of Our Lives: Memory in Infancy and Beyond*. Psychology Press. https://doi.org/10.4324/9781315785226

Bauer, P. J., Dikmen, S. S., Heaton, R. K., Mungas, D., Slotkin, J., & Beaumont, J. L. (2013). III. NIH Toolbox Cognition Battery (cb): Measuring episodic memory. *Monographs of the Society for Research in Child Development*, *78*(4), 34–48. https://doi.org/10.1111/mono.12033

Bauer, P. J., Zelazo, P. D., Philip David Zelazo, & Zelazo, P. D. (2013). IX. NIH Toolbox Cognition Battery (CB): Summary, conclusions, and implications for cognitive development. *Monographs of The Society for Research in Child Development*, *78*(4), 133–146. https://doi.org/10.1111/mono.12039

Benedict, R. H. (1997). Brief visuospatial memory test–revised. *Psychological Assessment Resources*.

Bhaumik, A. K., Reader, J. M., Rahman-Filipiak, A., Kairys, A., Duda, B., Heidebrink, J. L., Hampstead, B. M., Paulson, H. L., & Giordani, B. (2020). Diagnostic differentiation by NIH Toolbox-Cognition (iPad) for Alzheimer's disease, mild cognitive impairment (MCI), and healthy control participants: Neuropsychology/computerized neuropsychological assessment. *Alzheimers & Dementia*, *16*. https://doi.org/10.1002/alz.046086

Brandt, Å., Brandt, Å., Jensen, M. P., Søberg, M. S., Andersen, S. D., & Sund, T. (2020). Information and communication technology-based assistive technology to compensate for impaired cognition in everyday life: A systematic review. *Disability and Rehabilitation: Assistive Technology*, *15*(7), 810–824. https://doi.org/10.1080/17483107.2020.1765032

Carlozzi, N. E., Tulsky, D. S., Kail, R. V., Beaumont, J. L., & Beaumont, J. L. (2013). VI. NIH Toolbox Cognition Battery (CB): Measuring processing speed. *Monographs of The Society for Research in Child Development*, *78*(4), 88–102. https://doi.org/10.1111/mono.12036

Cheng, Y. (2009). When Cognitive Diagnosis meets Computerized Adaptive Testing: CD-CAT. *Psychometrika*, *74*(4), 619–632. https://doi.org/10.1007/s11336-009-9123-2

Crane, P. K., Narasimhalu, K., Gibbons, L. E., Pedraza, O., Mehta, K. M., Tang, Y., Manly, J. J., Reed, B. R., & Mungas, D. M. (2008). Composite scores for executive function items: Demographic heterogeneity and relationships with quantitative magnetic resonance imaging. *Journal of the International Neuropsychological Society*, *14*(5), 746–759. https://doi.org/10.1017/S1355617708081162

Distefano, R., Fiat, A. E., Merrick, J. S., Slotkin, J., Zelazo, P. D., Carlson, S. M., & Masten, A. S. (2021). NIH Toolbox executive function measures with developmental extensions: Reliability and validity with preschoolers in emergency housing. *Child Neuropsychology*, *27*(6), 709–717. https://doi.org/10.1080/09297049.2021.1888905

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F.-X., Balota, D. A., Brysbaert, M., Carreiras, M., Ferrand, L., Ktori, M., Maria Ktori, Maria Ktori, Perea, M., Rastle, K., Sasburg, O., Yap, M. J., Ziegler, J. C., & Grainger, J. (2011). Smart

phone, smart science: How the use of smartphones can revolutionize research in cognitive science. *PLOS ONE*, *6*(9), 24974. https://doi.org/10.1371/journal.pone.0024974

Dunn, L. M., & Dunn, L. M. (n.d.). *Peabody Picture Vocabulary Test (PPVT)—Third Edition*. Retrieved August 15, 2022, from https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft15145-000

Finkelman, M., Kim, W., Weissman, A., & Cook, R. (2014). *Cognitive Diagnostic Models and Computerized Adaptive Testing: Two new item-selection methods that incorporate response times*. *2*(3), 59–76. https://doi.org/10.7333/jcat.v2i3.43

Fox, S., Brown, L. J. E., Antrobus, S., Brough, D., Drake, R., Jury, F., Leroi, I., Parry-Jones, A. R., & Machin, M. (2020). *A smartphone app for frequent monitoring of cognition: Agile software co-designed alongside people with Dementia (Preprint)*. https://doi.org/10.2196/preprints.24483

Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development*, *10*(4), 483–527. https://doi.org/10.1016/0885-2014(95)90024-1

Gantois, P., de Lima-Júnior, D., de Sousa Fortes, L., Batista, G. R., Nakamura, F. Y., & de Souza Fonseca, F. (2021). Mental fatigue from smartphone use reduces volume-load in resistance training: A Randomized, single-blinded cross-over study. *Perceptual and Motor Skills*, *128*(4), 315125211016233–315125211016233. https://doi.org/10.1177/00315125211016233

Gershon, R. C., Cella, D., Fox, N. A., Havlik, R. J., Hendrie, H. C., & Wagster, M. V. (2010). Assessment of neurological and behavioural function: The NIH Toolbox. *The Lancet Neurology*, *9*(2), 138–139.

Gershon, R. C., Slotkin, J., Manly, J. J., Blitz, D. L., Beaumont, J. L., Schnipke, D., Wallner-Allen, K., Golinkoff, R. M., Gleason, J. B., Hirsh-Pasek, K., Adams, M. J., & Weintraub, S. (2013). IV. NIH Toolbox Cognition Battery (cb): Measuring language (vocabulary comprehension and reading decoding). *Monographs of the Society for Research in Child Development*, *78*(4), 49–69. https://doi.org/10.1111/mono.12034

Heaton, R. K., Akshoomoff, N., Tulsky, D. S., Mungas, D. M., Weintraub, S., Dikmen, S., Beaumont, J. L., Casaletto, K. B., Conway, K. P., Slotkin, J., & Gershon, R. (2014). Reliability and validity of composite scores from the NIH Toolbox Cognition Battery in adults. *Journal of The International Neuropsychological Society*, *20*(6), 588–598. https://doi.org/10.1017/s1355617714000241

Holdnack, J. A., Holdnack, J., Tulsky, D. S., Brooks, B. L., Jerry Slotkin, Jerry Slotkin, Slotkin, J., Gershon, R., Heinemann, A. W., & Iverson, G. L. (2017). Interpreting patterns of low scores on the NIH Toolbox Cognition Battery. *Archives of Clinical Neuropsychology*, *32*(5), 574–584. https://doi.org/10.1093/arclin/acx032

Hori, K., Fukuhara, H., & Yamada, T. (2022). Item response theory and its applications in educational measurement Part I: Item response theory and its implementation in R. *WIREs Computational Statistics*, *14*(2), e1531. https://doi.org/10.1002/wics.1531

Huebner, A. (2010). An Overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research and Evaluation*, *15*(3), 3.

Iverson, G. L., & Karr, J. E. (2021). Improving the methodology for identifying mild cognitive impairment in intellectually high-functioning adults using the NIH Toolbox Cognition Battery. *Frontiers in Psychology*, *12*, 724888. https://doi.org/10.3389/fpsyg.2021.724888

Jiang, T., & Fang, H. (2020). The influence of user interface design on task performance and situation awareness in a 3-player diner's dilemma game. *PLOS ONE*, *15*(3), 1–14. https://doi.org/10.1371/journal.pone.0230387

Jones, D. R., Dallman, A., Harrop, C., Whitten, A., Pritchett, J., Lecavalier, L., Bodfish, J. W., & Boyd, B. A. (2022). Evaluating the feasibility of the NIH Toolbox Cognition Battery for autistic xhildren and adolescents. *Journal of Autism and Developmental Disorders*, *52*(2), 689–699. https://doi.org/10.1007/s10803-021-04965-2

Jusko, M. L., Raiker, J. S., Campez, M., Smith, J. N., Fosco, W. D., Horta, L., Little, K., Espinal, K., Sanchez, G., Mattfeld, A. T., Gnagy, E. M., Greiner, A. R., Coles, E. K., & Pelham, W. E. (2021). Evaluation of working memory deficits in children with ADHD using the NIH list sorting working memory task. *Child Neuropsychology*, *27*(5), 613–620. https://doi.org/10.1080/09297049.2021.1876014

Kairys, A., Daugherty, A. M., Kavcic, V., Shair, S., Persad, C., Heidebrink, J. L., Bhaumik, A. K., & Giordani, B. (2021). Laptop-administered NIH Toolbox and Cogstate Brief Battery in community-dwelling Black adults: Unexpected pattern of cognitive performance between MCI and healthy controls. *Journal of The International Neuropsychological Society*, 1–10. https://doi.org/10.1017/s135561772100028x

Kavanaugh, B. C., Cancilliere, M. K., Fryc, A., Tirrell, E., Oliveira, J., Oberman, L. M., Wexler, B. E., Carpenter, L. L., & Spirito, A. (2020). Measurement of executive functioning with the National Institute of Health Toolbox and the association to anxiety/depressive symptomatology in childhood/adolescence. *Child Neuropsychology*, *26*(6), 754–769. https://doi.org/10.1080/09297049.2019.1708295

Ke Chen, Chen, K., Lou, V. W. Q., & Lo, S. S. C. (2021). Exploring the acceptance of tablets usage for cognitive training among older people with cognitive impairments: A mixed-methods study. *Applied Ergonomics*, *93*, 103381–103381. https://doi.org/10.1016/j.apergo.2021.103381

Kranou-Economidou, D., Maria Kambanaros, & Kambanaros, M. (2020). Combining intermittent Theta Burst Stimulation (iTBS) with computerized working memory training to improve language abilities in chronic Aphasia: A pilot case study. *Aphasiology*, 1–25. https://doi.org/10.1080/02687038.2020.1852165

L. Marciano, Marciano, L., Schulz, P. J., A.L. Camerini, & Camerini, A.-L. (2021). How smartphone use becomes problematic: Application of the ALT-SR model to study the predicting role of personality traits. *Computers in Human Behavior*, *119*, 106731. https://doi.org/10.1016/j.chb.2021.106731

Lauren R Ott, Schantell, M., Madelyn P Willett, Hallie J Johnson, Jacob A Eastman, Hannah J Okelberry, Tony W Wilson, Brittany K Taylor, & Pamela E May. (2022). Construct validity of the NIH toolbox cognitive domains: A comparison with conventional neuropsychological assessments. *Neuropsychology (Journal)*. https://doi.org/10.1037/neu0000813

Lo, H.-C., Peng, B.-K., & Li, C.-C. (2018). User-centered design process to develop motor speech disorder treatment assistive tool. *Advances in Intelligent Systems and Computing*, 52–58. https://doi.org/10.1007/978-3-319-96065-4_8

Lukin, V. N., Dzyubenko, A. L., & Chechikov, Yu. B. (2020). Approaches to user interface development. *Programming and Computer Software*, *46*(5), 316–323. https://doi.org/10.1134/s0361768820050059

Ma, Y., Carlsson, C. M., Wahoske, M. L., Blazel, H. M., Chappell, R. J., Johnson, S. C., Asthana, S., & Gleason, C. E. (2021). Latent factor structure and measurement invariance of the NIH Toolbox Cognition Battery in an Alzheimer's Disease research sample. *Journal of the International Neuropsychological Society*, *27*(5), 412–425. https://doi.org/10.1017/S1355617720000922

MacAulay, R. K., Boeve, A., & Halpin, A. (2021). Comparing psychometric properties of the NIH Toolbox Cognition Battery to gold-standard measures in socioeconomically diverse older adults. *Archives of Clinical Neuropsychology*. https://doi.org/10.1093/arclin/acab018

Marques-Costa, C., Pinho, M. S., Simões, M. R., & Prieto, G. (2020). The NIH Toolbox Cognition Battery App in healthy aging and cognitive decline due to Alzheimer's Disease and Mild Cognitive Impairment. *European Journal of Public Health*, *30*. https://doi.org/10.1093/eurpub/ckaa040.032

Martin, S. (2014). Measuring cognitive load and cognition: Metrics for technology-enhanced learning. *Educational Research and Evaluation*, *20*, 592–621. https://doi.org/10.1080/13803611.2014.997140

McGlohen, M., & Chang, H. H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, *40*(3), 808–821. https://doi.org/10.3758/brm.40.3.808

Mungas, D. M., Widaman, K. F., Zelazo, P. D., Tulsky, D. S., Heaton, R. K., Slotkin, J., Blitz, D. L., & Gershon, R. (2013). VII. NIH Toolbox Cognition Battery (CB): Factor structure for 3 to 15 year olds. *Monographs of The Society for Research in Child Development*, *78*(4), 103–118. https://doi.org/10.1111/mono.12037

Mungas, D., Reed, B. R., Crane, P. K., Haan, M. N., & González, H. (2004). Spanish and English Neuropsychological Assessment Scales (SENAS): Further development and psychometric characteristics. *Psychological Assessment*, *16*(4), 347–359. https://doi.org/10.1037/1040-3590.16.4.347

Neumann, D., Peterson, E. R., Underwood, L., Morton, S. M. B., & Waldie, K. E. (2021). Exploring the factor structure of the NIH Toolbox Cognition Battery in a large sample of eight-year-old children in Aotearoa, New Zealand. *Journal of The International Neuropsychological Society*, *27*(8), 825–834. https://doi.org/10.1017/s1355617720001265

Nielsen, J. (1999). *Designing Web Usability: A Book by Jakob Nielsen* (1st ed.). New Riders Publishing. https://www.nngroup.com/books/designing-web-usability/

Norman, D. (2013). *The Design of Everyday Things* (Revised and Expanded). Basic Books. https://mitpress.mit.edu/9780262640374/the-design-of-everyday-things/

Rao, S. M., Losinski, G., Mourany, L., Schindler, D., Mamone, B., Mamone, B., Reece, C., Kemeny, D., Narayanan, S., Miller, D. M., Bethoux, F., Bermel, R. A., Richard A. Rudick, Rudick, R. A., Richard A. Rudick, & Alberts, J. L. (2017). Processing speed test: Validation of a self-administered, iPad-based tool for screening cognitive dysfunction in a clinic setting. *Multiple Sclerosis Journal*, *23*(14), 1929–1937. https://doi.org/10.1177/1352458516688955

Rey, A. (1958). *L'examen clinique en psychologie. [The clinical examination in psychology]* (p. 222). Presses Universitaries De France.

Rudolph, S., Heisserman, J., & Culley, S. (2011). Design computing and cognition. *Ai Edam Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, *26*(2), 105–106. https://doi.org/10.1017/s0890060412000017

Rueda, M. R., Fan, J., McCandliss, B. D., Halparin, J. D., Gruber, D. B., Lercari, L. P., & Posner, M. I. (2004). Development of attentional networks in childhood. *Neuropsychologia*, *42*(8), 1029–1040. https://doi.org/10.1016/j.neuropsychologia.2003.12.012

Ruet, A., Deloire, M., Charré-Morin, J., Hamel, D., & Brochet, B. (2013). A new computerised cognitive test for the detection of information processing speed impairment in multiple sclerosis. *Multiple Sclerosis Journal*, *19*(12), 1665–1672. https://doi.org/10.1177/1352458513480251

Salthouse, T. A., Babcock, R. L., & Shaw, R. J. (1991). Effects of adult age on structural and operational capacities in working memory. *Psychology and Aging*, *6*(1), 118–127. https://doi.org/10.1037/0882-7974.6.1.118

Shao, Y., Mang, J., Li, P., Wang, J., Jie Wang, Deng, T., & Xu, Z. (2015). Computer-based cognitive programs for improvement of memory, processing speed and executive gunction during age-related cognitive decline: A meta-analysis. *PLOS ONE*, *10*(6). https://doi.org/10.1371/journal.pone.0130831

Shields, R. H., Kaat, A. J., Aaron J. Kaat, McKenzie, F. J., Drayton, A., Sansone, S. M., Coleman, J., Michalak, C., Claire Michalak, Claire Michalak, Riley, K., Karen Riley, Berry-Kravis, E., Gershon, R., Widaman, K. F., & Hessl, D. R. (2020). Validation of the NIH Toolbox Cognitive Battery in intellectual disability. *Neurology*, *94*(12). https://doi.org/10.1212/wnl.0000000000009131

Shih, S.-C., Kuo, B.-C., & Lee, S.-J. (2019). An online game-based computational estimation assessment combining cognitive diagnostic model and strategy analysis. *Educational Psychology*, *39*(10), 1255–1277. https://doi.org/10.1080/01443410.2018.1501468

Solomon, M., Andrew Gordon, Gordon, A., Iosif, A.-M., Geddert, R., Raphael Geddert, Krug, M. K., Mundy, P. C., & Hessl, D. R. (2020). Using the NIH Toolbox to assess cognition in adolescents and young adults with Autism spectrum disorders. *Autism Research*, *14*(3), 500–511. https://doi.org/10.1002/aur.2399

Tatiana Karpouzian-Rogers, Beth Makowski-Woidan, Alan Kuang, Hui Zhang, Angela Fought, Janessa Engelmeyer, M.-Marsel Mesulam, Sandra Weintraub, & Emily J Rogalski. (2021). NIH toolbox episodic memory measure differentiates older adults with exceptional memory capacity from those with average memory. *Alzheimers & Dementia*, *17*(S6). https://doi.org/10.1002/alz.057442

Taylor, B. K., Frenzel, M. R., Eastman, J. A., Wiesman, A. I., Wang, Y.-P., Calhoun, V. D., Stephen, J. M., & Wilson, T. W. (2020). Reliability of the NIH toolbox cognitive battery in children and adolescents: A three-year longitudinal examination. *Psychological Medicine*, 1–10. https://doi.org/10.1017/s0033291720003487

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. https://doi.org/10.1037/1082-989x.11.3.287

Throuvala, M. A., Pontes, H. M., Halley M. Pontes, Tsaousis, I., Griffiths, M. D., Rennoldson, M., & Kuss, D. J. (2021). Exploring the dimensions of smartphone distraction: Development, validation, measurement invariance, and latent mean differences of the

smartphone distraction scale (SDS). *Frontiers in Psychiatry*, *12*, 642634–642634. https://doi.org/10.3389/fpsyt.2021.642634

Tulsky, D. S., Carlozzi, N. E., Chevalier, N., Espy, K. A., Beaumont, J. L., & Mungas, D. (2013). V. NIH Toolbox Cognition Battery (CB): Measuring working memory. *Monographs of the Society for Research in Child Development*, *78*(4), 70–87. https://doi.org/10.1111/mono.12035

Uncapher, M. R., Thieu, M. K., & Wagner, A. D. (2016). Media multitasking and memory: Differences in working memory and long-term memory. *Psychonomic Bulletin & Review*, *23*(2), 483–490. https://doi.org/10.3758/s13423-015-0907-3

Vie, J.-J., Jill-Jênn Vie, Popineau, F., Bruillard, E., & Bourda, Y. (2017). *A review of recent advances in adaptive assessment*. *94*, 113–142. https://doi.org/10.1007/978-3-319-52977-6_4

Wang, Q., Sa Yang, Liu, M., Cao, Z., & Ma, Q. (2014). An eye-tracking study of website complexity from cognitive load perspective. *Decis. Support Syst.*, *62*, 1–10. https://doi.org/10.1016/j.dss.2014.02.007

Warsaw, R. E., Jones, A. R., Jones, A. M., Rose, A. K., Newton-Fenner, A., Alshukri, S., & Gage, S. H. (2021). Mobile technology use and its association with executive functioning in healthy young adults: A systematic review. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.643542

Weintraub, S., Bauer, P. J., Zelazo, P. D., Wallner-Allen, K., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Slotkin, J., Blitz, D. L., Carlozzi, N. E., Havlik, R. J., Beaumont, J. L., Mungas, D., Manly, J. J., Borosh, B. G., Nowinski, C. J., & Gershon, R. C. (2013). I. NIH Toolbox Cognition Battery (CB): Introduction and pediatric data. *Monographs of*

the Society for Research in Child Development, *78*(4), 1–15. https://doi.org/10.1111/mono.12031

Westerberg, H., Jacobaeus, H., Hirvikoski, T., Clevberger, P., Östensson, M.-L., Bartfai, A., & Klingberg, T. (2007). Computerized working memory training after stroke: A pilot study. *Brain Injury*, *21*(1), 21–29. https://doi.org/10.1080/02699050601148726

Wilkinson, G. S., & Robertson, G. J. (2006). Wide range achievement test (WRAT4). *Psychological Assessment Resources*.

Wilmer, H. H., Sherman, L. E., & Chein, J. (2017). Smartphones and cognition: A review of research exploring the links between mobile technology habits and cognitive functioning. *Frontiers in Psychology*, *8*, 605–605. https://doi.org/10.3389/fpsyg.2017.00605

Woodcock, R. W. (1997). The Woodcock-Johnson tests of cognitive ability: Revised. In *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 230–246). The Guilford Press.

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature Protocols*, *1*(1), 297.

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., Conway, K. P., Gershon, R., & Weintraub, S. (2014). II. NIH Toolbox Cognition Battery (CB): Validation of executive function measures in adults. *Journal of The International Neuropsychological Society*, *20*(6), 620–629. https://doi.org/10.1017/s1355617714000472

Appendix

Table 1

Cognitive Assessments in the NIHTB-CB

Process Assessed	Traditional Assessment	Test Description	NIHTB-CB Equivalent Test	Notes on Initial Validation
Cognitive Flexibility	Dimensional Change Cart Sort (DCCS) Frye, Zelazo, & Palfai, 1995; Zelazo, 2006	Two target cards are shown to the participants, who are then asked to sort subsequent cards by one dimension. Then, they are asked to switch and sort by another dimension using the same cards.	Toolbox DCCS Test Four blocks of images presented on a touchscreen - Practice, Preswitch, Postswitch, and Mixed. Participants are asked to match a central image with one of the laterally presented images on either side by colour or shape (white rabbit, green rabbit, white boat, green boat).	Excellent developmental sensitivity across childhood and test-retest validity. Adequate convergent validity among 8–15-year-olds for the Flanker Task. For younger children, correlations were higher for the DCCS than for the Flanker Task.
Inhibitory Control and Attention	Adapted Erisksen Flanker Task from the Attention Network Task Rueda et al., 2004	Participants indicate whether a center stimulus is oriented to the right or left while potentially surrounded by incongruently oriented stimuli (fish or arrows). They are presented with a combination of congruent and incongruent examples.	Toolbox Flanker Inhibitory Control and Attention Test For children, this test includes a practice block of fish stimuli, a fish block, and an arrows block. Participants are asked to press one of two onscreen buttons that corresponded with the direction of the middle stimuli. The word “middle” or an auditory reminder is available for younger children.	Correlations between the NIH Toolbox measures, and receptive vocabulary, declined with age, dropping by 43% for the Toolbox DCCS and 24% for the Toolbox Flanker as expected.
Episodic Memory	Imitation-based tasks (elicited and deferred imitation) Bauer, 2005, 2006, 2007	Novel toys are used to create a sequence that the infant imitates.	NIH TPSMT - Toolbox Picture Sequence Memory Test. Pictures with no inherent order (expected scenes as opposed to steps in a process) from three themes (Working on the farm, playing at the park, and going to the fair) are presented sequentially and audibly described. Then, the entire sequence is shown. The images are then randomized and available for the participant to select and move into the correct sequence. Number of pictures in the sequence increased with age.	Strong association between TPSMT scores and age for episodic learning and memory. Moderate correlation of the TPSMT with the combined validation measurement of memory in ages 3 - 15 , but higher than the correlation with the measure of vocabulary (supported the hypothesis). Subsequent norming and changes such as switching from a touchscreen to a mouse were mentioned as being necessary.
	Peabody Picture Test-4th Edition (PPVT-IV) Dunn & Dunn, 2007	Participant is presented with four images. The examiner says a word and asks the participant to point to the image that best represents the word. Participants are presented with words and letters, out of	Toolbox Picture Vocabulary Test or TPVT (New) Single words are presented auditorily, along with four images of objects, actions, and/or concepts.	Accuracy above other tests with a testing time of five minutes. Reliability can be assessed at the participant level.

Process Assessed	Traditional Assessment	Test Description	NIHTB-CB Equivalent Test	Notes on Initial Validation
		context, in decreasing familiarity and increasing phonological complexity, and asked to read the letter or work aloud.	Participant asked to select the image that is the best match for the auditory word until the completed a set number of items or were incorrect ten consecutive times.	Compared against the PPVT-IV, NIHTB-CB scores were marginally weaker.
	Wide Range Achievement Test Version 4- Reading Subtest (WRAT-IV) Wilkinson & Robertson, 2006	Participants must reproduce six geometric figures from memory in the correct location.		A wider examination of language may be a better predictor of future abilities and outcomes.
	Brief Visuospatial Memory Test-Revised (BVMT-R Total Recall) Benedict, 1997			
Episodic Memory (cont.)	Rey Auditory Verbal Learning Test (RAVLT) Rey, 1958	Participant is asked to repeat back as many words as they can remember after being read a list of 15 words by the examiner.	Toolbox Oral Reading Recognition Test or TORRT (New) Participant is shown a letter or word on the screen and asked to read it aloud, which is documented as correct or incorrect.	
Working Memory	Spanish and English Neuropsychological Assessment Scales (SENAS) Working Memory Task Crane et al., 2008; Mungas, Reed, Tomaszewski Farias, & DeCarli, 2005	Participants are shown an object, given the name of the object auditorily, and asked to place them in sequence.	Toolbox List Sorting Working Memory Test - TLSWMT (New) Participants are presented with images and the auditory names. They are asked to remember the stimuli and repeat them verbally according to their size. The lists gradually increase in size. The task starts with stimuli in one category and progresses to two categories, requiring the participant to track both categories.	Age effects were noted. The correlation of working memory to general functioning were somewhat lower with the TLSWMT for age 8 to 15. Limitations: Small sample sizes. Unable to distinguish between components of the phonological loop or visuospatial sketchpad. Norming and sensitivity were listed as future research.
Processing Speed	Salthouse's Pattern Comparison Task Salthouse, Babcock, & Shaw, 1991 Woodcock–Johnson Tests of Cognitive Ability	Participants respond with a specific button when presented with a specific stimulus. (Ex. Upper case = Blue button, Lower case = Red button) Participants are given 30 rows with a target geometric figure followed by 19 similar figures.	Toolbox Pattern Comparison Processing Speed Test Participants are presented with two patterns, and they must decide if they are the same or not the same.	Processing speed showed age-related performance patterns across development. Moderately similar results as traditional measurement tools, but convergent validity was lower than researchers anticipated. Touchscreen introduced some response variability (participant

Process Assessed	Traditional Assessment	Test Description	NIHTB-CB Equivalent Test	Notes on Initial Validation
	Woodcock, R.W., 1997	The participant must mark the 5 that match the initial target.		hesitated or lowered hand to the table between trials, resulting in a need to switch to keyboard use). When combined with Toolbox DCCS RT, and Toolbox Flanker Inhibitory Control and Attention Test RT findings, validation measures were higher.

(Akshoomoff et al., 2013; Bauer, Dikmen, et al., 2013; Bauer, Zelazo, et al., 2013; Carlozzi et al., 2013; Gershon et al., 2013; Mungas et al., 2013; Tulsky et al., 2013; Weintraub et al., 2013; Zelazo et al., 2013)

Table 2

Composite Crystalized, Fluid, and Overall Cognition Scores

Composite Score	Test Scores Included
Toolbox Crystallized Cognition Composite	Picture Vocabulary Test Oral Reading Recognition Test
Toolbox Fluid Cognition Composite	Dimensional Change Card Sort (DCCS) Test Flanker Inhibitory Control and Attention Test Picture Sequence Memory Test List Sorting Working Memory Test Pattern Comparison Processing Speed Test
Toolbox Cognitive Function Composite	Toolbox Crystallized Cognition Composite Toolbox Fluid Cognition Composite

(Akshoomoff et al., 2013)