Psychometric Properties of a Novel, Remotely Administered Visual N-Back Task in a Pilot Sample

Master's Programme in Psychology
Master's thesis

Author:
Joni Hytinkoski

Supervisor(s):
Dr. Kati Peltonen, University of Helsinki
Dr. Erika Jääskeläinen, University of Oulu

8.5.2025
Helsinki

Faculty: Faculty of Medicine
Degree Programme: Master's Programme in Psychology

Author: Joni Hytinkoski
Title: Psychometric Properties of a Novel, Remotely Administered Visual N-Back Task in a Pilot Sample
Level: Master's thesis
Month and year: May 2025
Number of pages: 56
Keywords: N-back, working memory, high-frequency testing, validation, pilot study
Supervisors: Kati Peltonen, Erika Jääskeläinen
Where deposited: Helsinki University Library (eThesis)
Abstract: *Aims:* Use of n-back tasks in high-frequency testing is a novel and promising paradigm. The present study aimed to comprehensively assess validity, reliability, feasibility and effects of age and training on CognitionKit N-Back in a Finnish pilot sample.

Methods: Participants (16 female, 1 male) included 17 adults (M_{age} = 48.47, SD_{age} = 12.00) with mild insomnia symptoms. Participants completed four week-long N-Back testing blocks, each separated by 3 months. Participants also completed comprehensive cognitive assessments at baseline and at a 6-month follow-up. Construct validity was assessed using multiple linear regression analyses and by analysing zero-order correlations between N-Back, age and clinical cognitive measures. Significant predictors and training effects were assessed using linear mixed models. Reliability was assessed using Pearson correlations and ICCs. Feasibility was assessed by analysing participant compliance.

Results: N-Back accuracy correlated strongly only with the expected predictors, i.e. age and other clinical measures of WM. Strongest predictors of N-Back accuracy were Spatial Span and participant age. Test-retest reliability of the task was high and comparable to other cognitive measures in common clinical use. Training affected performance over time. Participant compliance was excellent.

Conclusions: Most hypotheses received considerable support in this small sample. Per-participant data was remarkably comprehensive with excellent participant compliance displayed. Adequate reliability and validity were demonstrated. As expected, age also affected performance. Repetition improved performance over time, and training effects need to be considered when interpreting results for clinical use. The results of the present pilot suggest that clinical use may be deemed feasible.

Tekijä: Joni Hytinkoski
Työn nimi: Uuden etänä suoritettavan visuaalisen N-Back -testimenetelmän psykometriset ominaisuudet pilottiotoksessa
Työn laji: Maisterintutkielma
Kuukausi ja vuosi: Toukokuu 2025
Sivumäärä: 56
Avainsanat: N-back, työmuisti, jatkuva seurantamittaus, validointi, pilottitutkimus
Ohjaajat: Kati Peltonen, Erika Jääskeläinen
Säilytyspaikka: Helsingin yliopiston kirjasto (eThesis)
Tiivistelmä: *Tavoitteet:* N-back-menetelmien käyttö jatkuvassa seuranta-mittauksessa on uusi ja lupaava paradigma. Tämän tutkimuksen tavoitteena oli tarkastella CognitionKit N-Back-menetelmän validiteettia, reliabiliteettia ja käyttökelpoisuutta sekä iän ja harjoittelun vaikutusta suomalaisessa pilottiotoksessa.

Menetelmät: Tutkimukseen osallistui 17 suomalaista (16 naista, 1 mies) lievästä unettomuudesta kärsivää aikuista ($M_{ikä}$ = 48.47, $SD_{ikä}$ = 12.00). Osallistujat suorittivat neljä viikon mittaista N-Back-testauslohkoa kolmen kuukauden välein. Tutkimuksen alussa ja kuuden kuukauden seurantakäynnillä toteutettiin kattavat kognitiiviset tutkimukset. Rakennevaliditeettia tarkasteltiin lineaarisia regressio-analyysejä hyödyntäen sekä N-Backin, iän ja kliinisten mittareiden korrelaatioita tarkastellen. Merkittäviä ennustajia ja harjoittelun vaikutusta analysoitiin lineaarisin sekamallein. Reliabiliteetin tarkasteluun käytettiin ICC:tä ja Pearsonin korrelaatioita. Käyttökelpoisuutta tarkasteltiin analysoimalla suoritettujen testien määrää.

Tulokset: Tarkkuus N-Backissa korreloi vahvasti ainoastaan odotettujen ennustajien eli iän ja muiden kliinisten työmuistitestien kanssa. Vahvimmin tarkkuutta ennustivat ikä ja suoriutuminen Visuaalisissa sarjoissa. N-Backin toistomittaus-reliabiliteetti oli korkea ja vertautui muihin kliinisiin mittareihin. Harjoittelun määrä vaikutti suoriutumiseen. Osallistujat suorittivat pääosin riittävän määrän testejä.

Johtopäätökset: Suurin osa hypoteeseista sai tukea tässä otoksessa. Jokaisen osallistujan aineisto oli varsin kattava. Sitoutuneisuus ohjeistettuihin testimääriin ja aikatauluihin oli korkeaa. Reliabiliteetti ja validiteetti havaittiin hyväksyttäviksi. Ikä ja harjoittelu vaikuttivat suoriutumiseen odotetusti. Harjoitteluvaikutus tulee ottaa huomioon menetelmän tulosten kliinisessä tulkinnassa. Tämän pilottitutkimuksen tulosten perusteella menetelmä voidaan katsoa kliinisesti käyttökelpoiseksi.

Table of Contents

1 Introduction

1.1 Working Memory

Working memory (WM) as a construct refers to the limited capacity to temporarily store and manipulate mental information (Baddeley, 1992) especially while under distraction (Engle et al., 1999). WM can also be conceptualised as the ability to direct and keep attention on a limited number of task-relevant mental representations (Engle et al., 1999). WM capacity is considered to be an important substrate for successful execution of mental tasks such as reasoning, learning, comprehension and planning (Baddeley, 1992; Gevins & Smith, 2000) and is significantly correlated with general cognitive ability and fluid intelligence (Kyllonen and Christal, 1990; Perry et al., 2001). WM is notably limited, though the exact qualities of its limitations are a point of contention between psychologists (Cowan, 2012). Most commonly WM is considered to be limited in the number of meaningful units of data that can be actively maintained in WM, the amount of time these units can be held on to, or by the amount, type and intensity of interference that WM can tolerate (Cowan, 2012). WM function is generally measured using psychometric tests that require both storage and manipulation of mental information, often under additional cognitive load and processing demands (Haatveit et al., 2010). The tests most often used in measuring WM differ between the various subfields of psychology. In recent years, however, tasks originally borne of the experimental literature – such as the n-back, which has traditionally seen use mostly in neuroimaging studies (Redick & Lindsey, 2013) and neuroscience research (Schmiedek et al., 2014) – have gradually become more popular within the clinical setting (Haatveit et al., 2010). The present study aimed to evaluate a specific n-back task variant by assessing its psychometric properties, i.e., reliability and validity, and comparing it to relevant measures in common clinical use.

1.2 N-back

The n-back task is psychometric measure often conceptualised as an executive working memory (WM) task (Kane et al., 2007). The task has been widely used especially within cognitive neuroscience research for a number of decades (Kane et al., 2007), yet despite its popularity the detailed psychometric properties of the task are still somewhat unclear in the literature (Frost et al., 2021). According to Jaeggi et

al. (2010), the n-back task involves multiple different processes relevant to WM. The task demands the participant to match the present stimulus to the one n positions back in the presentation order. As such, monitoring, maintaining and updating the contents of working memory and encoding of each incoming stimulus are all required. As such, the task has agreeable face validity (Gajewski et al., 2018) as a WM measure requiring inhibition, interference resolution, selection and decision processes (Jaeggi et al., 2010; Jonides et al., 1997). As compared with e.g., simple span tasks better conceptualised as short-term memory (STM) measures, the n-back task has been considered a WM task due to the simultaneous requirements of processing, storing and updating of short-term memory content (Conway et al., 2005; Jonides et al., 1997; Kane & Engle, 2002) – a good match for a common, theoretical conceptualisation of working memory itself (i.e., Baddeley, 1992; Miyake & Shah, 1999). Many different versions of the task exist, both in the visuospatial and the verbal and written modalities. The n-back task has also been found to correlate with general fluid intelligence, executive function and attentional tasks (e.g., Kane et al., 2007; Gajewski et al., 2018; though see Haatveit et al., 2010 for contrasting views on the matter).

1.3 Psychometric Properties of the N-back Task

Despite presumed strengths of the task, the reliability and construct validity of n-back as a WM measure have recently come under some scrutiny (Gajewski et al., 2018). Middling reliabilities (e.g., Jaeggi et al., 2010) and intercorrelations (e.g., Frost et al., 2021; Miller et al., 2009; Kane et al., 2007) between n-back and other WM tasks have often been found in the literature, though many studies disagree with these findings (e.g., Haatveit et al., 2010; McMillan et al., 2007).

1.3.1 Construct Validity of the N-back Task

Validity of a psychometric measure is defined as the ability of a test to measure what it is claimed to measure (Anastasi, 1954; Colliver et al., 2012). Construct validity, then, refers to a specific measure's position within and connections with other interrelated measures in a network structure (Colliver et al., 2012), reflecting its ability to represent psychometric properties that are not de facto directly measurable. Construct validity is commonly divided into convergent and discriminant validity. For a measure to be deemed as displaying good convergent validity, significant

intercorrelations between the measure and other conceptually interrelated measures would need to be demonstrated (Campbell & Fiske, 1959). Discriminant validity, then, would be indicated by the measure not correlating with measures that are conceptually unrelated (Campbell & Fiske, 1959).

N-back, Working Memory and Short-Term Memory. Puzzlingly, some studies have suggested n-back performance to correlate quite strongly with performance in simple span tasks (e.g., Jaeggi et al., 2010), which are conceptualised as STM measures. Contrastingly, n-back performance has often been found to correlate less strongly with other common WM measures such as complex span tasks, e.g. operation span (Kane et al., 2007; Unsworth, 2010), counting span, math span, category span (Roberts & Gibson, 2002), reading span (Roberts & Gibson, 2002; Unsworth, 2010), symmetry span (Unsworth, 2010) and digit span backward (Miller et al., 2009). Correlations between n-back and complex span tasks in the aforementioned studies ranged from $r = .10$ to $r = .24$, which can be considered quite weak for tasks ostensibly measuring the same construct. As noted by Gajewski et al., (2018) correlations between n-back and simple STM tasks have been found to range between $r = .12$ to $r = .53$ in past studies (e.g., Colom et al., 2008; Gevins and Smith, 2000; Oberauer, 2005; Roberts and Gibson, 2002; Shelton et al., 2009).

Worthy of note in the literature are also the somewhat mixed findings regarding n-back's correlations with digit span forward and digit span backward. Digit span forward is usually considered to measure STM performance, while digit span backward is commonly conceptualised as a WM measure. Some studies (e.g., Jaeggi et al., 2010; Miller et al., 2009) have found stronger correlations between n-back and digit span forward, while some studies (e.g., Gajewski et al., 2018) have noted stronger correlations with digit span backward. Haatveit et al. (2010) also found n-back accuracy to significantly correlate with WAIS-IV Digit Span Backward at $r = .36$ and WAIS-IV Letter-Number Sequencing at $r = .33$ and found them to form a common component in a principal components analysis in healthy participants, adding to the often conflicting and contrasting findings in the literature.

Training transfer effects using n-back have also been demonstrated between the trained versions and other, untrained versions of the n-back task (e.g., Li et al., 2008). Most studies (e.g., Soveri et al., 2017), however, have suggested these effects to be mostly very small near-transfer effects confined to the class of n-back tasks,

with improved n-back performance not affecting performance in other WM measures (e.g., Redick & Lindsey, 2013). If n-back training truly improved general WM (as opposed to task specific) performance, similar improvements would be expected to be detectable in other WM tasks.

As noted by Gajewski et al. (2018), comparisons using composite outcome measures have offered a contrasting picture, however. For example, Shamosh et al. (2008) found n-back performance to correlate with a composite scale composed of rotation span, symmetry span, operation span and reading span scores at r = .55. Shelton et al. (2007, 2009) found their composite n-back score composed of multiple task versions to correlate with operation span at r = .46. Schmiedek et al. (2014) also found correlations ranging from r = .31 to r = .69 in younger and r = .27 to r = .66 in older participants between both spatial and numerical n-back performance and multiple other working memory measures, including a variety of span tasks. Many of the larger correlations between n-back and WM tasks in the study were similar in magnitude to the intercorrelations between other WM tasks. The writers postulated both n-back and complex span tasks as valid yet differing measures of working memory. Schmiedek, Hildebrandt et al. (2009) also found a latent WM variable including a variant of the n-back task and multiple other WM updating tasks to correlate with a latent WM variable composed of three different complex span tasks at r = .96. This finding was criticised by Redick & Lindsey in their 2013 meta-analysis, however, as rotation span was the only complex span task to correlate with n-back at the observed variable level, and the complex span factor loading for rotation span was much stronger than for the other two tasks. Both n-back and rotation span also shared their modality in common, as both were visuospatial tasks. Some promising results in the literature notwithstanding, Redick and Lindsey (2013) found n-back and complex span tasks to have an average meta-analytic correlation of only r = .20.

N-back and General Fluid Intelligence. STM and WM measures have generally been found to correlate quite strongly with Gf (e.g., Colom et al., 2008; Unsworth, 2010). Correlations between Gf and n-back performance would thus be expected if n-back truly is a measure of WM. Engle et al. (1999), indeed, found n-back performance to be linked with fluid intelligence but not with STM span task performance. Kane et al. (2007) found both n-back and operation span to correlate with general fluid

intelligence, but also that they correlated only weakly with each other. A factor analytic study by Burgoyne et al. (2024) found a WM factor composed of two n-back tasks to correlate with a complex span WM factor at r = .45 and with a general fluid intelligence (Gf) at r = .84. Other studies have also found n-back performance to correlate with Gf (e.g., Gevins and Smith, 2000; Hockey and Geffen, 2004). Contrastingly, other studies suggest no significant relationship between Gf and n-back performance (e.g., Haatveit et al., 2010). Unsworth (2010) also attempted to comprehensively assess the links between WM and Gf using structural equation modelling, though the included n-back task had to be discarded from all analysed models due to its weak correlations with other memory tasks. As such, n-back's correlations with Gf were unable to be assessed in the study.

N-back and Executive Function. Some studies have also indicated n-back performance to correlate with performance in Stroop tasks (e.g., Kwong See and Ryan, 1995), a paradigm considered to be a measure of executive function. Ciesielski et al. (2006), for example, found a variant of the n-back task to correlate with Stroop performance at r = .55, though Friedman et al. (2006), found n-back to correlate with Stroop only weakly (r = .10). Miller et al. (2009) found Stroop colour naming performance to correlate with two different n-back variants at r = .43 and at r = .46, though the correlations were not statistically insignificant. N-back was found to correlate significantly with Trail Making Test A, but not with WAIS-III Digit Span Backward, which is conceptualised as a measure of WM, nor WAIS-III Digit Span Forward, which is considered a STM task (Miller et al., 2009).

Findings on the construct validity of n-back remain somewhat inconclusive both due to the limited breadth and depth of the relevant literature and the mixed findings within. So far, even high-quality studies have provided contrasting findings with varying estimates of the magnitudes of relevant correlations.

1.3.2 Reliability of the N-back Task

Reliability of a psychometric measure is defined as the property of its consistency. A task can be considered reliable if and only if it is able to reproduce similar results under similar conditions when assessing the same data (Hammersley, 1987; Goode & Hatt, 1952, as cited in Hammersley, 1987). Different studies have produced extremely varying estimates for the reliability of the n-back measure, ranging from r = .09 to r =

.95 (e.g., Jaeggi et al., 2010; Burgoyne et al., 2024), depending on the specific formulation of reliability assessed. Most of the few reliability studies in the literature have assessed split-half reliability instead of temporally distanced test-retest reliability. Split-half findings in the literature have trended toward agreeable to good reliability (e.g., Kane et al., 2007; Burgoyne et al, 2024), though some studies disagree (e.g., Jaeggi et al, 2010). For high-frequency testing with repeated-measures computer-administered tasks where inter-rater variation is non-existent, however, test-retest reliability is the most relevant of the various subtypes of psychometric reliability. Of the few identified studies in the literature, Hockey and Geffen (2004) found test-retest reliability of the n-back task to be at least moderate. A prior 2022 feasibility study by Cormack et al. (2022, as cited in Cormack et al., 2024) found the test-retest reliability of a visual n-back task to be .80 when test data was aggregated over several measurements in a high-frequency testing paradigm similar to the present study. Due to the small number of relevant studies, test-retest reliability of the task yet remains somewhat of an open question.

1.3.3 N-back and Aging

Gajewski et al., 2018 found both performance accuracy and speed as well as the correlations between n-back and other measures to vary as a function of participant age. N-back accuracy was found to correlate with WAIS-IV Digit Span Backward and Forward only in middle-aged participants (aged 41–60), with these correlations being insignificant for younger (aged 20–40) and older (aged 41–60) participants. In addition to WAIS-IV Digit Span, n-back accuracy correlated with WAIS-IV Digit Symbol, Stroop 1 and 2 and TMT-B for these middle-aged participants. N-back accuracy correlated most strongly with d2, TMT-A, TMT-B, and Stroop 1–3 for younger participants, and with TMT-A, TMT-B, WAIS-IV Digit Symbol, d2, Word Fluency Test and Verbal Learning and Memory Test (VLMT) for older participants. In line with results from studies on other WM measures, n-back performance was found to decrease with age, both in speed and accuracy. Gajewski et al. (2018) found the assessed 2-back variant of the n-back task to be best conceptualised as a “complex cognitive task that measures a conglomerate of distinct cognitive functions that are differently involved depending on age”. Age has also been found to affect the efficiency of training, with older participants’ performance in the n-back task improving less over repetitions as compared to younger adults (Salminen et al.,

2016). A 2020 meta-analysis by Bopp & Verhaeghen also confirmed both n-back performance accuracy and speed to decline as a function of age, with generally larger age-related differences found when assessing more difficult versions of the task. Task response false positive rate has also been noted to increase with age (Schmiedek, Li et al., 2009). Findings in the literature regarding the specific effects of age on n-back performance are still somewhat mixed, though the majority of studies indicate declining performance and training efficiency with increasing age. Due to the small number of studies, further assessment of effects of age on n-back performance and training efficiency seems pertinent.

1.3.4 Training Effects, Feasibility and Utility of N-back in Clinical Populations

Training effects (i.e., effects of repetition improving task performance) have also been demonstrated in the literature (e.g., Li et al., 2008; Soveri et al., 2017). The existence of major training effects may make interpretation of results more complicated in high-frequency testing paradigms used for continuous monitoring of cognitive function. As such, the existence and significance of training effects is important to assess.

N-back performance accuracy has been demonstrated to be useful in discriminating between participants with schizophrenia-related cognitive decline and controls by Haatveit et al. (2010). Of the measures assessed in the study, n-back was also found to be the least influenced by Gf, education, gender and age. A 2018 study by Fraga et al. also found indications of n-back's potential utility in distinguishing between mild cognitive impairment and Alzheimer's Disease. Utility of the task in in detecting cognitive deficits caused by Parkinson's disease and in discriminating between patients with PD and controls was also postulated by Miller et al. (2009).

Due to the postulated clinical utility of the task, feasibility studies on high-frequency testing and cognitive monitoring using n-back paradigms have been completed, albeit in limited numbers. A prior 2022 feasibility study by Cormack et al. (2022) found participant compliance to be acceptable, although not excellent. In the study, poor compliance was defined as a participant completing less than 50% of their scheduled self-testing sessions and was observed in 22% of participants. Worthy of note was also their finding that 61.5% participants completed 100% or more self-test sessions.

Due to the small number of relevant feasibility studies, feasibility of the task in high-frequency cognitive monitoring remains an important research question.

1.4 Motivations for The Present Study

The varying findings regarding the psychometric properties of n-back and its correlations with other WM tasks paint a complicated picture. In contrast with the wide usage of the task, the literature on the specific psychometric properties of the n-back paradigm yet remains comparatively sparse. The theoretical underpinnings of what specific facets of WM is n-back actually measuring also remain somewhat unclear. Questions regarding the task's construct validity, reliability, and the propriety of its classification as a pure WM task remain open. Further study of the psychometric properties of the n-back task and its variants thus seems pertinent. If the task is to be considered a useful WM measure in a clinical environment, at least agreeable correlations with established, clinical WM tasks should be expected.

Most of the relevant studies in the literature have focused on single-session n-back assessments in a laboratory environment. The present study aimed to assess the construct validity and reliability of Cambridge Cognition's (2023) CognitionKit 2-Back variant N-Back task (hereafter referred to as *N-Back*, italicized and with first letters capitalised, better facilitating differentiation between the assessed task and the n-back paradigm) in high-frequency testing. CognitionKit *N-Back* is a remote visuospatial task specifically designed for high-frequency testing. The present study was conducted in a Finnish pilot sample and aimed to assess whether the remotely administered *N-Back* test is psychometrically adequate and feasible for use in continuous monitoring of WM function. In patient cases where progressive deficits are a part of the prognosis of the condition – such as Alzheimer's disease – remote high-frequency testing with an instrument like the CognitionKit *N-Back* would be considerably less invasive and disruptive to patients' lives than repeated visits at the hospital or the clinician's office, while also likely being much more affordable.

1.5 Research Questions and Hypotheses

The present study aimed to comprehensively assess relevant properties of the CognitionKit *N-Back* task for clinical high-frequency monitoring of WM function. The research questions of the present study concerned the task's psychometric

properties (i.e., construct validity and reliability), age-related effects on performance, training effects and feasibility of high-frequency testing with the CognitionKit *N-Back* task.

Construct validity of the n-back paradigm yet remains an open question, and thus the first research question (RQ1) was "What is the construct validity of the CognitionKit *N-Back* task?". First, it was hypothesised that (H1.1) significant and acceptably high correlations between *N-Back* performance accuracy and other, relevant clinical WM measures (i.e., WMS-III Spatial Span Backward, WAIS-IV Digit Span Backward, and WAIS-IV Digit Span Sequencing) would be found, thus displaying adequate convergent validity. The strongest correlates and their possible interactions with training effects were to be assessed in detail. Second, in line with previous findings in the literature it was hypothesised that (H1.2) the *N-Back* performance accuracy would correlate less strongly – albeit still significantly – with tasks generally conceptualised as measures of executive function (i.e., Stroop, TMT-B and WAIS IV-Coding). Third, it was hypothesised that (H1.3) *N-Back* performance accuracy would not correlate significantly with other assessed cognitive tasks, thus displaying adequate discriminant validity.

As the estimates for reliabilities of n-back tasks in the literature are mixed and as the high-frequency testing paradigm for continuous monitoring of cognitive function is a novel application for n-back, the second research question (RQ2) was "What is the test-retest reliability of the CognitionKit *N-Back* task?". In line with findings of Cormack et al. (2022) it was hypothesised that (H2) *N-Back* would display adequate test-retest reliability when each participant's performance scores are aggregated within week-long testing blocks, averaging out day-to-day variability in performance.

The literature on effects of age on n-back performance in high-frequency testing paradigms yet remains considerably small. As such, the third research question (RQ3) was "Does age effect CognitionKit *N-Back* performance accuracy?" In line with the literature on age effects on performance in other WM tasks, it was hypothesised that (H3) participant age would negatively and independently affect performance accuracy in the *N-Back* task even in the presence of other, significant predictor variables. Additionally, an interaction between age and effects of training was presumed possible and thus was to be assessed if an independent effect of age was found.

Learning and training related improvements in a task may bring additional difficulties in interpretation of results in high-frequency testing paradigms used for continuous monitoring of cognitive function. Training effects in different n-back tasks have been demonstrated in the literature, and as such the fourth research question (RQ4) was “Does performance accuracy in the CognitionKit *N-Back* task improve with training?”. It was hypothesised that (H4) independent effects of training (i.e., improved performance over time due to increased repetitions) would be clearly observable even in the presence of other, possibly more significant predictor variables. Interactions between training and age, and training and other significant predictor variables were presumed possible. These interactions were to be assessed if an independent effect of training on performance was found.

Lastly, the present study aimed to assess the real-world feasibility of high-frequency testing with the *N-Back* task by assessing participant testing compliance and patterns of missing data. As such, the fifth research question (RQ5) was “Is participant compliance with the CognitionKit *N-Back* high-frequency testing paradigm adequate?” In line with previous results by Cormack et al. (2022) it was hypothesised that (H5) participant compliance would be satisfactory. Furthermore, it was expected that the number of missing data points would be limited to the extent as to not cause significant issues with interpretation of results.

2 Methods

2.1 Ethical Statement

The present study was conducted according to the Declaration of Helsinki, and the study received an ethical approval from the regional ethics committee of the Northern Ostrobothnia Hospital District.

2.2 Participants

Participants included 17 (16 female, 1 male) healthy Finnish adults (M_{age} = 48.47, SD_{age} = 12.00) from a pilot subsample of participants recruited for the APSY Oulu – "The Effects of Antipsychotics in Insomnia, Anxiety and Depression" study (Jääskeläinen et al., 2021) conducted at the University of Oulu. The writer of the present study was one of the data collectors. All participants were Finnish-speaking working-age adults with mild insomnia symptoms and belonged to the control group of the APSY study. None of the participants had recently started any new medications affecting wakefulness, fatigue or the central nervous system. Therapeutic equilibrium at time of assessment was a requirement for inclusion in the pilot sample. Most of the control participants were recruited from the staff of the University of Oulu. Participation was completely voluntary, and participants were not monetarily compensated. The data was pseudonymised after collection. Written, informed consent was provided by each participant.

2.3 Materials

2.3.1 WMS-III

Immediate auditory memory: Logical Memory I & Verbal Paired Associates I. These subtests are used as measures of immediate auditory memory (Wechsler, 1997, 2008b). In Logical Memory I (LM I) the participant is verbally presented with two stories and tasked with carefully listening to and remembering their contents. After presentation of each story the participant is asked to immediately recall and verbally repeat the story as accurately as they can. The first story is presented once to the participant, while the second story and its associated immediate free recall task is presented twice. The Logical Memory subtest is considered to be a measure of episodic memory (Ahn et al., 2019), assessing verbal comprehension and

memorisation of semantically meaningful narrative information (Groth-Marnat & Wright, 2016, as cited in Tyni, 2022). The task has been found to be sensitive in detecting even subtle changes in memory performance associated with mild cognitive impairment as well useful in discriminating between different types of dementia (Ahn et al., 2019). The scores attainable in each story-task range from 0 to 25, with higher scores indicating better performance. The participant is awarded credit for each meaningful detail recalled, and the scores thus achieved across the three trials are summed. This sum score is used as the outcome measure. In Verbal Paired Associates I (VPA I) the participant is tasked with memorising and immediately recalling eight semantically unrelated word pairs over three sequential trials. The possible scores for each trial range from zero to eight, with higher scores indicating better performance. The task is used to assess explicit episodic memory and verbal learning performance, and it is quite sensitive to diverse deficits in memory function (Uttl et al., 2002). The primary outcome measure – and the one chosen for the present study – is the sum score of all three trials, with a separate learning score computable by assessing the differences in performance between the trials. Together, these subtests comprise the WMS-III Auditory Immediate memory index.

Delayed auditory memory: Logical Memory II & Verbal Paired Associates II. These subtests are used as measures of delayed auditory memory (Wechsler, 1997, 2008b). In Logical Memory II (LM II) the participant is tasked with recalling and verbally repeating the stories presented in LM I. This task is presented with a delay of 30 minutes after completion of LM I and is used to assess delayed free recall narrative memory performance. The scores attainable for each of the two story-tasks range from 0 to 25. The participant is awarded credit for each meaningful detail recalled, with the summed score over both trials used as the outcome measure. In Verbal Paired Associates II (VPA II) the participant is sequentially presented with the first words of the word pairs learned in VPA I after a delay of 30 minutes. The participant is tasked with recalling and repeating the other halves of the word pairs. The task is used as a non-narrative verbal measure of cued, delayed recall. The number of word pairs (zero to eight) correctly recalled is used as the outcome measure. Verbal Paired Associates II and Logical Memory II together constitute the WMS-III Auditory Delayed memory index.

Spatial Span Forward & Backward. The Spatial Span subtest (Wechsler, 1997, 2008b) is used as a measure of visual short-term and working memory. It has been suggested to function as a visual analogue of the of the Wechsler Digit Span subtest (The Psychological Corporation, 1997, as cited in Wilde & Strauss, 2002), though this suggestion is somewhat contested in the literature (see e.g., Wilde et al., 2004). The Forward and Backward tasks are commonly taken to measure different functions (Kaplan et al., 1999, as cited in Wilde & Strauss, 2002). Spatial Span Forward is commonly considered to be a measure of visual short-term memory. Spatial Span Backward is often conceptualised as a WM measure due to the increased processing demands of the task (Tulsky, 2004), though some studies suggest that both tasks actually measure mostly the same construct (see e.g., Tulsky, 2004; Wilde & Strauss, 2002; Wilde et al., 2004). Nonetheless, the task is in common clinical use and has been shown to be sensitive to working memory deficits associated with e.g., schizophrenia (Manglam et al., 2010) and Alzheimer's disease (Kessels et al., 2015). In the Spatial Span Forward task, a stimulus board – on top of which lay 10 cubes in semi-random visual order – is placed between the examiner and the participant. The participant is first asked to attend to a series of taps the examiner presents using the board and is then tasked with repeating the tapping sequence themselves. The tapping sequences are presented in an ascending order of difficulty, with the length of the sequences increasing by one after each two trials. The sequence lengths vary from two to nine. In the Spatial Span Backward task, the participant is tasked with repeating the presented sequences in reverse order. The lengths of the sequences again vary from two to nine, with sequence lengths increasing by one every two trials. Each task is discontinued when the participant is unsuccessful in correctly repeating neither sequence of a specific length. The number of correctly repeated sequences is used as the outcome measure for each task, with possible scores in each ranging from 0 to 16. Together, the two tasks comprise the Spatial Span subtest of the WMS-III Working Memory index.

2.3.2 WAIS-IV

Similarities. The Similarities (Wechsler, 2008a, 2012) subtest is a test of abstract verbal reasoning that requires semantic processing and making inferences about concepts and categories. The participant is verbally presented with two stimulus words at a time and is then tasked with describing semantic similarities between the

word pairs (e.g., “in what way are an apple and an orange similar to each other?”). The scoring of each trial is affected by the quality of the answer, with a maximum of two points awarded for a correct answer of sufficient quality. The word pairs are presented in an ascending order of difficulty, and the task is discontinued when the participant gives three incorrect answers in a row. The sum score (0 to 36) attained by the participant is used as the outcome measure. The task is a core subtest of the WAIS-IV Verbal Comprehension Index.

Block Design. Block Design (Wechsler, 2008a, 2012) is a task that is principally used as a measure of visuospatial construction ability (Joung et al., 2021) and spatial visualisation (Yang et al., 2014), requiring both spatial reasoning skills and fine motor skills. In this time limited test, the participant is tasked with replicating visual patterns using either four or nine multicoloured cube-shaped blocks depending on the item. The stimulus designs are presented as two dimensional shapes on a piece of paper. Scoring of the task is somewhat involved compared to most other WAIS-IV subtests, as the scores for items 9–14 are affected both 1) by whether the design was correctly replicated within the time limit and 2) by time spent on the item, with faster completion times resulting in higher scores. Scores for items 5–8 are awarded purely on successful completion of the item within the time limit. The exact time spent on each item is disregarded. Scoring on items 1–4 is affected both by 1) correct replication of the stimulus designs and 2) by the number of trials required to do so. The task is continued until the participant fails to correctly replicate the stimulus design for two sequential trials. The score (0 to 66) thusly achieved is the principal outcome measure of the task. A common alternative outcome measure for the task is a version of the score with no time bonus awarded, as it is less affected by modest deficits in fine motor skills and by slower, more careful problem-solving strategies less well-suited for timed tasks (e.g., Weiss et al., 2016). The timed score was chosen as the outcome measure for the present study due to its higher between-participants variability. The task is a core subtest of the WAIS-IV Perceptual Reasoning Index.

Digit Span Forward, Backward & Sequencing. Digit Span Forward (Wechsler, 2008a, 2012) is generally categorised as a test of short-term memory (STM) due to its relatively simple and uninvolved task demands (Gignac & Weiss, 2015) as compared to WM tasks. The task requires the participant to memorise and recall simple, random series of verbally presented single digits in the exact order of presentation.

The lengths of the digit series range from two to nine. After presentation of stimuli, the participant must verbally repeat the memorised digits. As compared with Digit Span Backward and Digit Span Sequencing, Digit Span Forward requires no further manipulation or processing of the relevant memory content. Digit Span Backward and Digit Span Sequencing (Wechsler, 2008a, 2012) are better categorised as complex span tasks (The Psychological Corporation, 2002, as cited in Wilde et al., 2004) and tests of WM capacity, as the tasks require mental manipulation of the order of the memorised stimuli (Gignac & Weiss, 2015). In Digit Span Backward the participant is again verbally presented with a random series of single digits to be memorised. After presentation of stimuli, the participant is asked to recall and verbally repeat the memorised digit series in the reverse order. The lengths of the digit series range from two to eight. In Digit Span Sequencing the participant is tasked with memorising, recalling and verbally repeating the presented digit series in ascending numerical order. The series lengths range from two to nine. For each of the three Digit Span tasks two separate trials per each series length are presented directly after each other, with ascending difficulty. Each task is continued until the participant is unsuccessful in correctly recalling neither digit series of a specific length. The scores attainable for each task range from 0 to 16, with higher scores indicating better performance. The principal outcome measure for the whole subtest is the sum score (0 to 48) of all three tasks combined. Due to its utility as a general measure of memory performance, the sum score was chosen as one of the predictor variables for the present study. A common alternate outcome measure specific to each task is the Span score, i.e., the length of the longest digit sequence accurately recalled on at least one of the two trials of a specific series. The maximum Span score attainable thus ranges from zero to nine for Digit Span Forward and Sequencing, and zero to eight for Digit Span Backward. These per-task Span scores were also assessed as separate predictor variables due to the aforementioned differences in their relations with STM and WM. Together, the three tasks constitute the Digit Span core subtest of the WAIS-IV Working Memory Index.

Coding. In the Coding task (Wechsler, 2008a, 2012) the participant is supplied with a pen and a paper task sheet and is tasked with filling in multiple rows of empty boxes with symbols that correspond to nine possible digits presented one by one above the boxes. The corresponding symbols are presented paired with the nine digits in a key positioned at the top of the task sheet. The test is timed, with an absolute time limit of

120 seconds. Full completion of all symbol-digit pairings on the sheet within the allotted time is rare. The number of correct digit-symbol pairings achieved within the time limit is thus used as the principal outcome measure, with the score attainable ranging from 0 to 135. The task is mainly considered to be a measure of processing speed, while also requiring attention, cognitive flexibility and sufficient motor coordination (Ryan et al., 2015). The task is a core subtest of the WAIS-IV Processing Speed Index.

2.3.3 TMT-A & TMT-B

The Trail Making Test (Poutiainen et al., 2010) originally published in the Army Individual Test Battery (1944) is a timed neuropsychological test primarily used to assess participants' abilities for visual search and scanning, processing speed and mental flexibility (Tombaugh, 2004), though motor speed also affects time spent on both tasks (Arbuthnott & Frank, 2000). TMT-A and TMT-B are some of the most widely used neuropsychological measures (Tombaugh, 2004) and are often used as somewhat general measures of executive function, as performance on the task is quite sensitive to multiple different neurological disabilities and deficits (Mitrushina et al., 1999, as cited in Tombaugh, 2004), such as Alzheimer's disease (Amieva et al., 1998) and Parkinson's disease (Olchik et al., 2017). In TMT-A the participant is supplied with a pen and is tasked with drawing lines on a sheet of paper to connect 25 encircled numbers in a sequence (i.e., 1–2–3–4, etc.). In TMT-B the requirements of the task are quite similar to TMT-A, though in TMT-B the participant has to connect 13 encircled numbers and 12 letters in an alternating pattern (i.e., 1–A–2–B, etc.). For both TMT-A and TMT-B the primary outcome measures are the time elapsed on task completion as well as the TMT-B / TMT-A completion time ratio. TMT-B scores thus also reflect set-switching performance as a measure of executive control, with the TMT-B / TMT-A completion time ratio providing additional information independent of the effects of motor coordination fluency and visual scanning speed (Arbuthnott & Frank, 2000). The number and quality of task errors provide supplementary qualitative information and insight on the participant's tendency to favour either completion speed or accuracy. Completion times for each condition were chosen as outcome measures for the present study due to their generality and comparative simplicity.

2.3.4 Stroop

The Stroop task (Stroop, 1935) is a test of cognitive flexibility (Gluud et al., 2019) and selective attention (Goldberg & Bougakov, 2005) that measures set-switching fluency and ability to suppress habitual responding in favour of a less familiar, more deliberate one in the presence of conflicting cues (Stenberg et al., 2016). Performance on the Stroop task is sensitive to executive function deficits, with e.g. longer task completion times found in patients with schizophrenia due to increased interference effects in the colour incongruence-condition (Henik & Salo, 2004). The Finnish version (Nybo et al., 2000) used for the present study first tasks the participant with rapidly reading aloud 100 colour words (“red”, “blue”, “yellow” or “green”) presented sequentially on a sheet of paper. In the second condition the participant is tasked with rapidly naming the colours of 100 separate stimuli, sequentially presented as coloured series of the letter x. In the last condition the participant is presented with 100 colour words printed in an incongruent colour (e.g., the word “yellow” printed in blue ink) and is tasked with rapidly naming the colours of the ink each word is printed in, without getting distracted by the conflicting semantic information. Condition completion times, differences in completion times between conditions and number of errors are some the most common outcome measures used, though practices vary widely between and within different research traditions and countries (Scarpina & Tagini, 2017). Completion times and the numbers of errors were used as the outcome measures for the present study due to their simplicity and widespread clinical utility.

2.3.5 Finger Tapping

The Finger Tapping task is a test commonly used as a simple measure of psychomotor speed (Mendez, 2021). In the version of the test (Reinvall & Poutiainen, 2008) used in the present study, the participant is tasked with tapping a button on the Tapping-device with their thumb as many times as possible within a ten second time limit. The task is completed once with each hand, starting with the dominant hand. The outcome measure is the number of taps achieved within the time limit, counted separately for the dominant and the non-dominant hand.

2.3.6 CognitionKit Visual *N-Back* (2-Back) Task

Visual n-back tasks such as the CognitionKit *N-Back* task (Cambridge Cognition, 2023) used in the present study are commonly conceptualised as visual working memory updating tasks (e.g., Jonides et al., 1997; Watter et al., 2001). In addition to the task's varying correlations with other WM tasks, n-back performance has also been shown to correlate with other measures of cognitive function, selective attention and task switching (Cormack et al., 2024; Espeland, 2013). N-back tasks demand the participant to continuously update the contents of their working memory while simultaneously preventing contamination by older, now-irrelevant information (Frost et al., 2021; Owen et al., 2005), as the previously presented stimuli change their roles from possible target stimuli to distractors (Frost et al., 2021; Watter et al., 2001). As such, the task has also been categorised as measuring executive working memory function (Haatveit et al., 2010).

In the assessed 2-Back version of the CognitionKit *N-Back* task, participants use their personal smartphones to view a sequentially presented stimulus stream of abstract shapes. Participants tap the screen to indicate a match between the presently shown stimulus and the stimulus presented two trials prior. The participants need to also suppress the tendency to respond to lure trials (i.e., stimuli matching the present stimulus, yet presented at locations n−3 or n−1). As tested, each CognitionKit *N-Back* assessment takes 45 seconds to complete and is composed of 7 target and 23 non-target trials, for a total of 30 trials. In the assessed, present version of the *N-Back* task, each stimulus is presented for 1 second, with an empty fixation grid shown to the participant between trials. See Figure 1 for a visual representation of the task.

Performance accuracy in the task is assessed using the d-prime measure (*d'),* an outcome variable commonly used in the relevant literature to assess n-back performance. The *d'* is a parametric measure of sensitivity measuring the accuracy of a participant's ability to discriminate between stop and go trials. The *d*' is calculated for each set of trials by subtracting the z-transformed false-alarm rate from the z-transformed hit rate (Macmillan & Creelman, 1990). A *d'* of 0 indicates accuracy of 50% on both target hits and target rejections (Haatveit et al., 2010). The maximum *d'* value – indicating perfect accuracy – for the assessed CognitionKit *N-Back* test is 3.23, with the minimum value being -3.30. The better the participant can minimise

false alarms and maximise correct hits, the higher their *d'* value is. *d'* was chosen as the outcome variable for *N-Back* performance in the present study due to previous suggestions of its reliability and relevance in the literature.

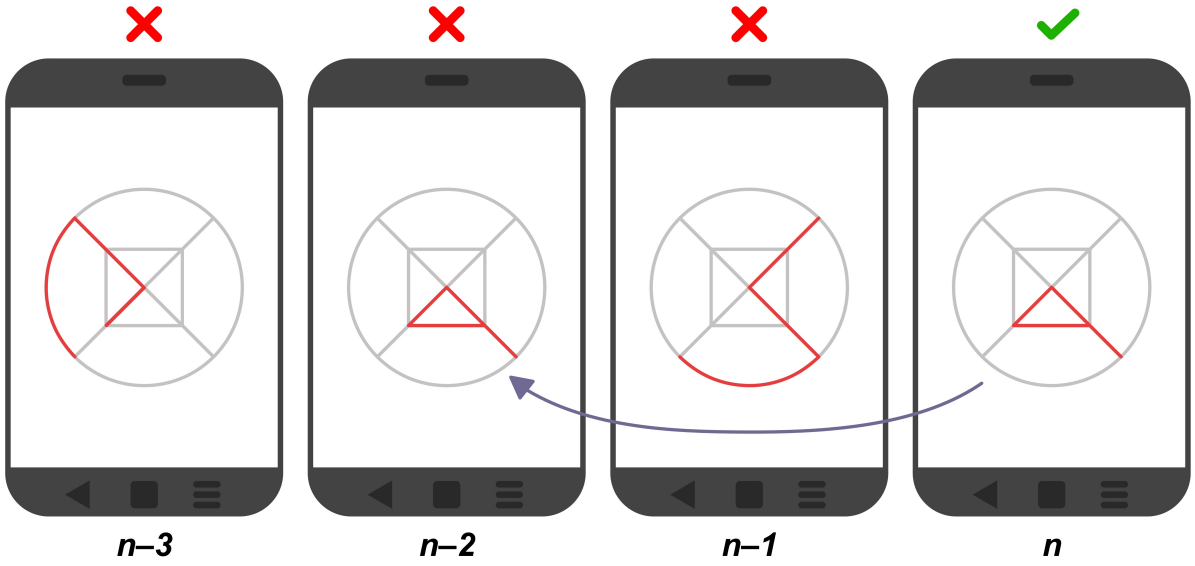

Figure 1. Illustration of the CognitionKit *N-Back* (2-Back) task. The participant is sequentially presented a series of visual stimuli on the screen of a smartphone. The participant indicates whether the current stimulus matches the stimulus presented two stimuli prior (n-2) by tapping the screen. The participant would be expected to tap the screen in the scenario pictured above.

2.4 Procedure

Each participant completed a comprehensive in-person cognitive assessment (i.e., WMS-III, WAIS-IV, TMT, Stroop, and Finger Tapping) at baseline and at a 6-month follow-up meeting. All assessments were performed in Finnish. After their in-person cognitive assessments each participant also completed a supervised *N-Back* practice trial. Participants were then instructed to independently complete the remote *N-Back* tasks using their personal smartphones every morning and evening during the subsequent week, at both baseline and at 6 months. In addition to these testing weeks, the participants were also requested to complete *N-Back* testing weeks at 3 months and 12 months post baseline, for a total of four week-long *N-Back* testing blocks. See Figure 2 for a visual representation of the testing schedule.

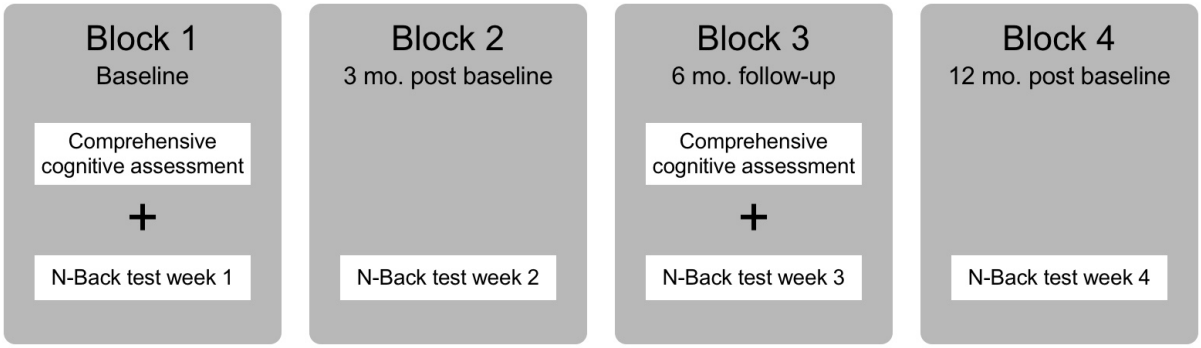

Figure 2. Participant testing schedule.

In addition to verbal instructions the participants were given an information sheet detailing the *N-Back* testing instructions and their planned schedule. The information sheet recommended the participants to complete their morning *N-Back* sessions after waking up and their evening sessions after 6 PM. On Mondays of each testing block the *N-Back* trials presented for each participant were an unscored practice variant, which was used to confirm that the participant understood the task instructions and that they were able to complete the *N-Back* task properly. The scored variant of the task was used for the rest of the trials, with the aim of gathering 12 scored *N-Back* trial data points per participant in each assessment block to facilitate analysis of high-frequency testing results. Most participants completed close to the instructed number of 12 repetitions per block (see Table 1), though fidelity to the instructed schedule varied both in number and timing both within and between participants.

Table 1. Number of valid *N-Back* test sessions completed by each participant.

ID	Block 1	Block 2	Block 3	Block 4	Total
1	12	9	11	12	44
2	9	7	-	-	16
3	12	8	12	12	44
4	12	7	11	13	43
5	11	9	11	-	31
6	12	-	-	12	24
8	8	-	10	8	26
9	7	5	6	-	18
10	9	44	10	10	73
11	9	10	10	-	29
12	12	-	-	-	12
13	12	-	12	-	24
14	11	-	-	-	11
15	12	9	11	-	32
17	11	10	10	-	31
18	10	33	8	-	51
20	12	12	9	-	33
Total	181	163	131	67	542
N	17	12	13	6	

2.5 Data Coding and Cleanup

The specific times of day each participant tended to complete their morning and evening test sessions varied considerably between and within participants. The data gathered by the CognitionKit test application did not specifically differentiate between morning and evening sessions, and thus completion times were coded manually. Due to visually apparent clustering, sessions completed between 04:00 AM and 04:59 PM were coded as morning tests, while tests completed between 05:00 PM and 03:59 AM were coded as evening tests. Assessment of possible differences between morning and evening scores was deemed important, as wakefulness and vigilance are known to affect performance in a wide range of psychometric tests.

The number of tests completed and the fidelity in following the planned *N-Back* testing schedule varied greatly between participants, which resulted in minor complications in categorising the test session data into the planned blocks. *N-Back* data for some participants was gathered temporally further away from the cognitive assessment dates than planned (e.g., with a participant completing their *N-Back* testing two weeks after their baseline cognitive assessment, instead of during the week immediately following it). For each participant, the first *N-Back* testing weeks (regardless of their possible incompleteness, i.e., a participant having completed only 9 out of 12 scheduled *N-Back* test sessions) conducted closest to baseline and the 6-month follow-up were coded as belonging to blocks 1 and 3 respectively. *N-Back* testing weeks completed within approximately a month following or preceding the respective cognitive assessment dates were deemed as valid for inclusion. Testing weeks completed closer to the planned 3-month and 12-month follow-up blocks were coded as belonging to blocks 2 and 4 respectively.

44 of the 737 recorded *N-Back* test sessions were removed due to being mere boots and thus containing no usable data. 131 sessions were removed due to being non-scored practice tests completed on Mondays. 2 recorded sessions were removed due to incorrectly being scored tests completed on a Monday instead of the mandated non-scored practice test. 18 sessions were removed due to incomplete data corresponding to an unfinished test. 542 test sessions across 17 participants remained for inclusion for analyses.

3 Results

3.1 Descriptive Statistics of *N-Back* Test Performance Accuracy

The present analyses of *N-Back* scores focused on the d-prime (*d'*) outcome variable, as is common in the relevant research literature. A total of 542 *N-Back d'* scores were judged viable for further analysis after cleaning the data. The primary blocks of interest were blocks 1 and 3, as their timing coincided with the baseline and 6-month follow-up in-person cognitive assessments. As such, the data from blocks 1 and 3 were used for most analyses. The complete dataset (i.e. scores from all four blocks) was used for assessing the effects of time of day, training and day-to-day variability in performance.

Block 1 included a total of 181 valid session scores across 17 participants (M = 10.647 valid sessions / participant, range = 7–12), while block 3 included a total of 131 valid session scores across 13 participants (M = 10.077 valid sessions / participant, range = 6–12). *N-Back* session *d'* scores below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR were defined as outliers. 14 per-participant statistical outliers in *N-Back d'* scores were identified. The outliers were decided to be left in for all analyses as per-participant data was quite robust and as one of the aims of the study was assessing the ecological validity and viability of the measure.

The *N-Back d'* score grand mean assessed across all four blocks was 1.73 (SD = 0.76), with scores ranging from -0.50 to 3.23 in a theoretically possible range of -3.30 to 3.23. The block 1 grand mean of *d'* scores was 1.56 (SD = 0.67) with scores ranging from -0.50 to 3.23. In block 3 the grand mean of *d'* scores was 1.86 (SD = 0.81) with a range of -0.15 to 3.23. No apparent ceiling effects were detected.

3.2 Preliminary Analyses

3.2.1 Time-of-Day Effects

A priori an effect of time of day on test scores was presumed possible, and thus possible time-of-day effects were assessed with a two-tailed paired-samples *t*-test. Each participant's morning and evening *N-Back d'* scores across all blocks were averaged resulting in two values for each participant (i.e., a personal morning mean *d'* score and a personal evening mean *d'* score). A visual assessment of the Q-Q plot

and a Shapiro-Wilk test indicated that the assumption of normality was not violated (W = 0.90, p = .070). The difference between *N-Back d'* morning scores (M = 1.75, SD = 0.59) and evening scores (M = 1.72, SD = 0.55) was nonsignificant, $t(16)$ = 0.65, p = .522. Both morning and evening *d'* scores were thus deemed fit for analysis as a single dataset for all further statistical procedures.

3.2.2 Differences in Per-Participant *N-Back* Accuracy Between Blocks

The difference in *N-Back d'* scores between blocks 1 and 3 was assessed with a two-tailed paired samples *t*-test to preliminarily assess possible training effects. Each participant's scores across all test sessions in block 1 and block 3 were averaged respectively for this test, resulting in two values per participant (i.e., a personal block 1 mean *d'* score and a personal block 3 *d'* mean score). A visual assessment of the Q-Q plot and a Shapiro-Wilk test indicated that the assumption of normality was not violated (W = 0.94, p = .420). A statistically significant difference ($t(12)$ = -3.96, p = .002) was found between participants' block 1 (M = 1.58, SD = 0.49) and block 3 (M = 1.90, SD = 0.65) personal mean *d'* scores. Lending preliminary support for hypothesis H4, this suggested the presence of training effects. Training effects were decided to be further analysed using linear mixed effects models. See Figure 3 for a visualisation of the difference in scores within and between participants for block 1 and 3.

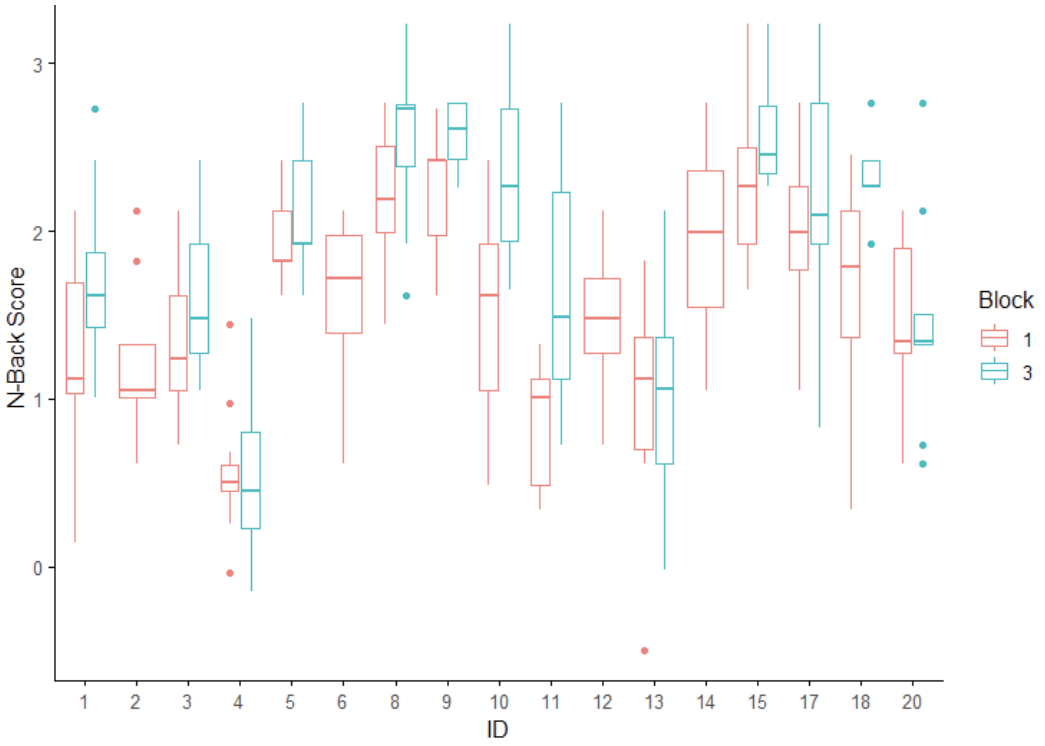

Figure 3. *N-Back d'* boxplots for each participant in blocks 1 and 3.

3.3 Construct Validity

3.3.1 Methodology for Assessing Construct Validity

The first primary research question (RQ1) was "What is the construct validity of the CognitionKit *N-Back* task?". Construct validity of the *N-Back d'* measure was preliminarily assessed using multiple linear regression analyses to examine correlations between *d'* and other the cognitive measures. Each participant's personal *d'* scores across all sessions in blocks 1 and block 3 were averaged respectively, resulting in two values for each participant (i.e., a personal block 1 mean *d'* score and a personal block 3 mean *d'* score) which were used as the dependent measures for these analyses. Zero-order intercorrelations for the cognitive measures that significantly correlated with *N-Back d'* in either block 1 or block 3 are presented in Tables 2 and 3. Though training effects may account for some of the differing correlations between blocks 1 and block 3, it was deemed important to compare *N-Back* performance with the most recent data available from the in-person cognitive assessments for each block.

Table 2. Zero-order intercorrelations of *d'* and other psychometric measures in block 1.

Measure	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.
1. *N-Back d'*																	
2. Age	**-.648****																
3. WMS-III Log. Mem. I, Story B2	.397	-.047															
4. WMS-III Log. Mem. I, Total	.453+	-.254	**.912*****														
5. WMS-III Spatial Span FW	.290	**-.513***	.196	.251													
6. WMS-III Spatial Span BW	**.642****	-.450+	.448+	**.548***	.366												
7. WMS-III Spatial Span Total	**.579***	**-.579***	.401	**.496***	**.796*****	**.854*****											
8. WMS-III Log. Mem. II, Story B	.383	-.136	**.863*****	**.949*****	.146	**.512***	.415+										
9. WMS-III Log. Mem. II, Total	.384	-.216	**.730*****	**.912*****	.197	.469+	.414+	**.909*****									
10. TMT-A Time	**-.753*****	**.667****	-.227	-.394	-.350	-.380	-.443+	-.306	-.462+								
11. TMT-A Errors	**.519***	**-.532***	.114	.046	.031	.169	.127	-.003	-.159	-.273							
12. TMT-B Time	**-.523***	.280	-.368	-.407	-.151	**-.498***	-.408	-.276	-.329	**.531***	-.370						
13. Stroop 1 Time	**-.498***	.194	-.445+	-.427+	-.058	-.423+	-.307	-.355	-.291	.484*	-.328	**.891*****					
14. Stroop 3 Time	**-.571***	**.582***	-.085	-.237	-.375	**-.493***	**-.530***	-.148	-.254	**.631****	-.326	**.689****	**.553***				
15. WAIS-IV Digit Span BW Span	.223	-.207	.218	.317	.354	.364	.435+	.332	.339	-.186	-.057	-.302	-.333	-.414+			
16. WAIS-IV Digit Span Seq. Span	**.652****	-.402	**.639****	**.679****	.446+	**.704****	**.707****	**.567***	**.603***	**-.560****	.109	**-.558***	**-.544***	-.428+	**.570***		
17. WAIS-IV Digit Span Total Raw	**.650****	-.416+	**.507***	**.542***	.436+	**.689****	**.692****	**.541***	**.493***	**-.494***	.234	-.464+	-.475+	-.423+	**.733*****	**.849*****	
18. WAIS-IV Coding	**.607****	**-.658****	.395	**.509***	.451+	**.541***	**.604***	.325	.386	**-.674****	.374	**-.733*****	**-.655****	**-.665****	.049	**.491***	.338

Note. Raw, unscaled scores were used for all measures to facilitate assessing correlation with participant age. Measures that had significant correlations with *d'* in block 1 are marked in bold. Measures that had significant correlations with *d'* in both blocks 1 and 3 are underlined. Measures that had no significant correlations with *d'* in either block 1 or block 3 are not included in the table.

*** = $p < .001$, ** = $p < .01$, * = $p < .05$, + = $p < .1$.

Table 3. Zero-order intercorrelations of *d'* and other psychometric measures in block 3.

Measure	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.
1. *N-Back d'*																	
2. Age	**-.763****																
3. WMS-III Log. Mem. I, Story B2	**.589***	**-.641****															
4. WMS-III Log. Mem. I, Total	**.596***	**-.583***	**.923*****														
5. WMS-III Spatial Span FW	**.616***	**-.587***	**.496***	**.537***													
6. WMS-III Spatial Span BW	**.854*****	**-.492***	.446+	**.579***	**.604***												
7. WMS-III Spatial Span Total	**.829*****	-.600*	**.524***	**.624****	**.883*****	**.907*****											
8. WMS-III Log. Mem. II, Story B	**.594***	**-.601***	**.815*****	**.856*****	.452+	.469+	.515*										
9. WMS-III Log. Mem. II, Total	**.594***	**-.562***	**.775*****	**.859*****	.430+	.539*	.544*	**.942*****	**.594***								
10. TMT-A Time	-.092	.388	-.258	-.301	-.158	-.238	-.223	**-.496***	-.375								
11. TMT-A Errors	.083	.167	-.196	-.132	.185	.066	.137	-.120	-.142	.336							
12. TMT-B Time	-.419	.352	**-.706****	**-.736*****	-.160	-.407	-.324	**-.741*****	**-.685****	.353	**.509***						
13. Stroop 1 Time	-.116	.319	**-.640****	**-.669****	.058	-.058	-.003	**-.807*****	**-.765*****	.424+	.367	**.745*****					
14. Stroop 3 Time	-.376	.540*	**-.763*****	**-.635****	**-.590***	-.164	-.408	**-.538***	-.439+	.176	.251	.480+	.387				
15. WAIS-IV Digit Span BW Span	**.792****	**-.522***	.455+	.497*	.271	**.718****	**.565***	.407	.373	-.387	-.155	**-.513***	-.224	-.245			
16. WAIS-IV Digit Span Seq. Span	.300	-.170	**.607****	**.536***	.103	.264	.209	**.492***	**.483***	-.040	-.374	**-.766*****	**-.571***	-.520*	.412		
17. WAIS-IV Digit Span Total Raw	**.612***	-.393	**.681****	**.636****	.281	.577*	.488*	**.524***	.445+	-.244	-.258	**-.743*****	-.431+	**-.509***	**.775*****	**.756*****	
18. WAIS-IV Coding	.543+	**-.613****	**.751*****	**.820*****	**.561***	.472+	**.574***	**.808*****	**.822*****	-.251	-.269	**-.701****	**-.682****	**-.681****	.390	**.658****	**.519***

Note. Raw, unscaled scores were used for all measures to facilitate assessing correlation with participant age. Measures that had significant correlations with *d'* in block 3 are marked in bold. Measures that had significant correlations with *d'* in both blocks 1 and 3 are underlined. Measures that had no significant correlations with *d'* in either block 1 or block 3 are not included in the table.

*** = $p < .001$, ** = $p < .01$, * = $p < .05$, + = $p < .1$.

3.3.2 Convergent Validity

It was hypothesised (H1.1) that CognitionKit *N-Back d'* would correlate strongly and significantly with other measures conceptualised as clinical WM tasks. As seen in Tables 2 and 3, the only measures that significantly correlated with *N-Back d'* at both timepoints were participant age, WMS-III Spatial Span Backward raw score, WMS-III Spatial Span total raw score, and WAIS-IV Digit Span total raw score. The strongest single predictor of *N-Back d'* in the present data was WMS-III Spatial Span Backward (r = .642, p < .01 at block 1, r =.854, p < .001 at block 3), a result rendered somewhat unsurprising due to similarities in modality and cognitive task demands. Interestingly, the correlation between *d'* and WAIS-IV Digit Span Backward was significant at block 3 but not block 1, while the correlation between *d'* and WAIS-IV Digit Span Sequencing was significant at block 1 but not block 3. Nonetheless, these results on the statistical significance and magnitude of the relevant correlations lend preliminary support for hypothesis H1.1, with CognitionKit *N-Back d'* displaying adequate convergent validity with other relevant WM measures.

3.3.3 Discriminant Validity

It was hypothesised (H1.2) that correlations between *N-Back d'* and tasks conceptualised as measuring executive function would be significant, albeit weaker than the correlations between *d'* and WM tasks. It was also hypothesised (H1.3) that other (i.e., non-WM and non-EF) tasks would not correlate significantly with *N-Back d'*. Indeed, *d'* scores did not correlate significantly with Digit Span Forward at either timepoint (r = .296, p > .1 at block 1, r = .300, p > .1 at block 3). Neither did *d'* scores correlate strongly or reliably with measures of executive function, attention or processing speed, as correlations between *d'* and Stroop, TMT and WAIS-IV Coding were significant at block 1 but insignificant at block 3. This inconsistency could – at least partially – be explained by the small sample size and participant dropout. The significant zero-order correlations between *d'* and WMS-III Logical Memory in block 3 were deemed likely spurious due to statistical outliers and the small sample size (n = 13) at block 3. Support for hypothesis H1.2 was somewhat unclear in the present sample due to inconsistent correlations between *N-Back d'* and executive function tasks. Hypothesis H1.3 was supported, however, with *N-Back d'* displaying adequate discriminant validity. *d'* scores achieved consistent statistically significant

correlations only with the more difficult span tasks commonly conceptualised as WM tasks.

3.4 Reliability

The second research question (RQ2) was "What is the test-retest reliability of the CognitionKit *N-Back* task?". In line with findings of Cormack et al. (2022), hypothesis H2 posited that *N-Back d'* would display adequate test-retest reliability when each participant's performance scores are aggregated within week-long testing blocks, averaging out day-to-day variability in performance. Test-retest reliability was operationalised as the Pearson correlation of scores between block 1 and block 3. *N-Back* data gathered in the continuous high-frequency testing paradigm for each block were pooled and averaged respectively into personal per-block mean *d'* scores for each participant. As can be seen in Table 4, test-retest reliability of the CognitionKit *N-Back d'* measure thusly processed is quite high and compares favourably to the test-retest reliabilities of other psychometric measures.

Table 4. Test-retest reliabilities of CognitionKit *N-Back d'* and other measures.

Measure	Test-retest Reliability
N-Back d'	.893***
WMS-III Logical Memory I, Story B2	.277
WMS-III Logical Memory I, Total Score	.597*
WMS-III Spatial Span Forward	.795***
WMS-III Spatial Span Backward	.755***
WMS-III Spatial Span Total	.864***
WMS-III Logical Memory II, Story B	.439+
WMS-III Logical Memory II, Total Score	.568*
TMT-A Time	.460+
TMT-A Errors	-.133
TMT-B Time	.784***
Stroop 1 Time	.760***
Stroop 3 Time	.851***
WAIS-IV Digit Span Backward	.279
WAIS-IV Digit Span Sequencing	.390
WAIS-IV Digit Span Total	.653**
WAIS-IV Coding	.921***

Note. Test-retest reliabilities between block 1 and block 3 were assessed using Pearson correlations. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

An intraclass correlation analysis produced an ICC of 0.504 for per-trial *N-Back* session *d'* scores clustered within participants. The ICCs varied between blocks, with per-participant *N-Back* session *d'* scores in block 1 having an ICC of 0.450, an ICC of 0.535 in block 2, an ICC of 0.561 in block 3 and an ICC of 0.740 in block 4. This variance in per-block ICCs was expected, as training could be expected to decrease within-participant variation in performance accuracy. Sample size also varied considerably between blocks and was smallest in block 4, as previously discussed and shown in Table 1. These results indicate considerable day-to-day variance in *N-Back* performance within participants, though the measure displayed adequate test-retest reliability when per-participant *d'* scores were aggregated over multiple sessions in a week-long testing block. The results were deemed to support hypothesis H2, with *N-Back d'* displaying adequate test-retest reliability with each participant's performance scores aggregated. Aggregation of *N-Back* test results over multiple sessions thus seems advisable to guarantee reliability of the measure in between-participant designs.

3.5 Assessment of Age and Other Predictor Variables of *d'* With Linear Multiple Regression Analyses

The third research question (RQ3) was "Does age effect CognitionKit *N-Back* performance accuracy?". It was hypothesised (H3) that participant age would negatively and independently affect *N-Back d'* even in the presence of other significant predictor variables. The aforementioned zero-order correlational analyses provided preliminary support for hypothesis H3, as age was found to be a significant predictor of *N-Back d'*.

Age and performance in related psychometric measures of WM were expected to both predict *N-Back d'* independently and to share common covariance with each other. I.e., age was expected to affect performance in both *N-Back* and other WM tasks, and performance in a single WM task was expected to correlate with performance in other tasks. As such, the intercorrelations had to be considered when analysing these predictor variables of *d'* further. To facilitate the study of interaction effects with age, untransformed raw scores with no age standardisation were used for all clinical psychometric measures. Linear multiple regression analyses were conducted on the data to assess the explanatory power of each significant predictor variable while

controlling for the effects of other significant predictor variables, to clarify both the effects of age and the general construct validity of the *N-Back* task.

A bidirectional stepwise linear regression was run to predict block 1 *N-Back d'* scores from candidate variables identified in the aforementioned zero-order correlational analyses, namely TMT-B time, Stroop 3 time, WAIS-IV Coding raw score, WAIS-IV Digit Span total raw score, WMS-III Spatial Span Backward raw score, WMS-III Spatial Span total raw score and participant age. At each step, variables to be included or discarded were chosen based on their contribution to the model's residual sum of squares (RSS), and the final model was selected according to minimum AIC. This resulted in a significant model, $F(2, 14) = 10.28$, $p = .002$, adjusted $R^2 = .537$. As the individual predictor variables were further examined, the analysis indicated that age ($t = -2.444$, $p = .028$) and WAIS-IV Digit Span total raw score ($t = 2.457$, $p = .028$) were significant predictors, with other variables discarded.

VIF analysis indicated a moderate level of collinearity for WMS-III Spatial Span total raw score (VIF = 6.25, 95% CI = [4.34, 9.25]), TMT-B time (VIF = 6.21, 95% CI = [4.31, 9.19]) and WAIS-IV Coding raw score (VIF = 6.80, 95% CI = [4.70, 10.09]). The analysis indicated that participant age (VIF = 3.55, 95% CI = [2.55, 5.18]), WAIS-IV Digit Span total raw score (VIF = 3.25, 95% CI = [2.36, 4.74]), WMS-III backward Spatial Span raw score (VIF = 4.40, 95% CI = [3.12, 6.46]) and Stroop 3 time (VIF = 3.04, 95% CI = [2.22, 4.41]) had low collinearity. Even though the VIF analysis did not bring forth alarming results regarding collinearity, many of the variables shared a noticeable amount of variance with each other, as can be seen in Tables 2 and 3. This indicates that the covariance between *d'* and other tasks conceptualised as WM measures is largely common and shared between the variables, resulting in multiple WM tasks not explaining much unique variance in *N-Back* accuracy. The assumptions of homoscedasticity and normality were assessed using the Breusch-Pagan test and the Shapiro-Wilk test respectively. Error variance appeared homoscedastic ($p = .535$), and residuals appeared normally distributed ($p = .273$).

A second bidirectional stepwise linear regression was also completed to predict block 3 *N-Back d'* scores from candidate variables identified as significant in the aforementioned zero-order correlational analysis and the previous stepwise linear regression, namely TMT-B time, Stroop 3 time, WAIS-IV Coding raw score, WAIS-IV Digit Span total raw score, WMS-III Spatial Span Backward raw score, WMS-III

Spatial Span total raw score and participant age. At each step, variables to be included or discarded were again chosen based on their contribution to the model's RSS. The final model was selected according to minimum AIC. This resulted in a significant model, $F(3, 9) = 21.11$, $p < .001$, adjusted $R^2 = .834$. Further examination of the individual predictor variables indicated that age ($t = -2.935$, $p = .017$) and WMS-III Spatial Span Backward raw score ($t = 3.252$, $p = .010$) were significant predictors, with WAIS-IV Digit Span total raw score as a statistically insignificant predictor ($t = 1.305$, $p = .224$) and with the rest of the candidate variables discarded.

WAIS-IV Digit Span total raw score being an insignificant predictor variable in this second analysis would on a first glance contrast the results of the block 1 multiple regression analysis. This may, however, be explained by the collinearity of the independent variables, as VIF analysis indicated a high level of collinearity for the WAIS-IV Digit Span total raw score (VIF = 12.44, 95% CI = [9.35, 16.69]), WMS-III Spatial Span Backward raw score (VIF = 32.39, 95% CI = [24.09, 43.68]), WMS-III Spatial Span total raw score (VIF = 23.65, 95% CI = [17.63, 31.86]) and Stroop 3 time (VIF = 10.94, 95% CI = [8.24, 14.66]). Moderate collinearity was indicated for TMT-B time (VIF = 7.92, 95% CI = [6.01, 10.58]) and WAIS-IV Coding raw score (VIF = 7.02, 95% CI = [5.34, 9.35]). The analysis indicated that participant age (VIF = 2.45, 95% CI = [1.96, 3.17]) had low collinearity. These results indicate strong collinearity between many of the variables. I.e., the variables shared in common a considerable amount of covariance with each other and thus would not be best conceptualised as truly independent. This can also be confirmed through a visual analysis of Tables 2 and 3. The assumptions of homoscedasticity and normality were assessed using the Breusch-Pagan test and the Shapiro-Wilk test respectively. Error variance appeared homoscedastic ($p = .189$), and residuals appeared normally distributed ($p = .918$). The results of both linear multiple regressions analyses lend considerable support to the assumption that the covariance *d'* shares with WM tasks is largely shared in common between the assessed variables, with multiple WM tasks not explaining much unique variance in *N-Back* accuracy.

As analyses indicated different models for blocks 1 and 3, the predictor models and variables were further assessed by cross testing the model indicated by block 1 on block 3 and the model indicated by block 3 on block 1. The models were compared using *p*-values, AICs and R^2 values. The significant model indicated for block 1

included age and WAIS-IV Digit Span total raw score as significant predictors of *d'* performance, while the model indicated for block 3 included age and WMS-III Spatial Span Backward raw score. For block 1, the model ($F(2, 14) = 10.28$, $p = .002$, adjusted $R^2 = .537$) including age ($t = -2.444$, $p = .028$) and WAIS-IV Digit span total raw score ($t = 2.457$, $p = .028$) had an AIC of 19.103. The model ($F(2, 14) = 9.43$, $p = .003$, adjusted $R^2 = .513$) including age ($t = -2.308$, $p = .037$) and WMS-III Spatial Span Backward raw score ($t = 2.246$, $p = .041$) had an AIC of 19.964. Both models were statistically significant. Accordingly, for block 3 the model ($F(2, 10) = 13.47$, $p = .001$, adjusted $R^2 = .675$) including age ($t = -3.620$, $p = .005$) and WAIS-IV Digit Span total raw score ($t = 2.326$, $p = .042$) had an AIC of 20.648. The model ($F(2, 10) = 28.78$, $p < .001$, adjusted $R^2 = .822$) including age ($t = -2.875$, $p = .017$) and WMS-III Spatial Span Backward raw score ($t = 4.264$, $p = .002$) had an AIC of 12.801. Both models were significant. In all assessed models both WAIS-IV Digit Span total raw score and WMS-III Spatial Span Backward raw score worked separately as significant predictors with participant age but became nonsignificant when both were included in a model together. Assessed as a whole, a model including age and WMS-III Spatial Span total raw score fit both blocks well, while the model including age and WAIS-IV Digit Span total raw score had a comparatively worse fit for block 3. The present results suggest that age and WM function separately predict *N-Back d'* performance, with the WMS-III Spatial Span Backward raw score being the most optimal predictor of *d'* of the assessed WM variables. This result is somewhat unsurprising due to the congruence of its modality with the visuospatial nature of the CognitionKit *N-Back* task. These results provided considerable additional support for hypotheses H1.1 and H1.3, and as such suggest adequate construct validity. Hypothesis H3 was also supported, as age remained a significant predictor of *d'* in all models.

3.6 Training Effects

As improvement through training in n-back tasks has been demonstrated in the literature, the fourth research question (RQ4) was "Does performance accuracy in the CognitionKit *N-Back* task improve with training?". It was hypothesised (H4) that independent effects of training would be clearly observable even in the presence of other, possibly more significant predictor variables. As preliminary analyses showed significant differences in aggregated per-participant *N-Back d'* performance between blocks, it was deemed important to further analyse training effects in the presence of

other covariates of *N-Back* performance in a repeated-measures paradigm. This was assessed by building multiple repeated measures linear mixed models on the data, with the personal *d'* score in each separate testing session acting as a dependent variable clustered within a specific participant. The effect of training was operationalised by using the per-participant ordinal number of each personal *N-Back* test session as an independent predictor variable for the *d'* value of that specific test session. Interactions between age and training, and training and other significant predictor variables were presumed possible. These interactions were to be assessed if an independent effect of training on performance was found.

All between-participant variables (i.e., participant age and personal scores in other psychometric measures) used values gathered in the comprehensive cognitive assessment completed by each participant in block 1. Block 1 values were deemed to be the most acceptable choice, due to risk of training and retest effects contaminating the predictive validity and power of the clinical psychometric tasks at block 3. The clinical psychometric measures have only been validated for non-repeated testing. To facilitate comparison of variables and the study of interaction effects with age, untransformed raw scores with no age standardisation were used for all clinical psychometric measures.

Five separate models were assessed in addition to the null model. See Table 5 for comparison of models. The null model indicated that 28.68% of the random effect variance was located in the between-participants level, with 28.22% of the random effect variance existing in the trial level. A model comparison using *p*-values and the minimum AIC and BIC method revealed model 4 (AIC = 855.68, BIC = 890.04) to be the optimal model of the ones assessed ($p = .035$). The model indicated that after controlling for the random effects of participant ($s^2 < .001$, $SD = 0.36$) and the interaction of training and participant ($s^2 < .001$, $SD = 0.01$), training had a statistically significant effect on personal *N-Back* session *d'* scores within participants ($t(17.99) = 4.71$, $p < .001$). The between-participants level independent variables of age ($t(21.07) = -2.519$, $p = .020$) and WMS-III Spatial Span Backward raw score ($t(21.10) = 2.339$, $p = .029$) were also significant predictors of between-participants differences in *N-Back d'* scores. Adding WAIS-IV Digit Span raw score into the model did not significantly enhance the model's explanatory power when WMS-III Spatial Span Backward raw score was already included. Training was a significant predictor

of *N-Back d'* session scores in all assessed models. These results indicate that training has an independent significant effect on *N-Back* performance accuracy even when effects of other significant predictor variables are included in the model, supporting hypothesis H4. Hypothesis H3 also received additional support, as age remained a significant predictor of *N-Back* performance in all models. Hypotheses H1.1 and H1.3 also garnered further support, as after age and training were considered, the strongest predictors of between-participant differences in *N-Back* performance were other WM measures.

Multicollinearity of the variables in model 4 was assessed by conducting a VIF analysis. The analysis revealed low collinearity for training (VIF = 1.02, 95% CI = 1.00 to 2.00), age (VIF = 1.02, 95% CI = 1.18 to 1.46) and WMS-III Spatial Span Backward raw score (VIF = 1.02, 95% CI = 1.17 to 1.43). Two further models were built to assess possible interaction effects between 1) training and age and 2) training and WMS-III Spatial Span Backward raw score. Both models revealed nonsignificant interactions, *p*s = .685 and .198.

Table 5. Comparison of the assessed linear mixed models and their fit indices.

Model	Dependent Variable		Fixed Effects		Random Effects	Fit Indices			LRT Test Against Nested		
						AIC	BIC	LL	*df*	χ^2	*p* (> χ^2)
Null	Session *d'* score	~	1	+	(1 | ID)	916.17	929.06	-455.09			
Model 1	Session *d'* score	~	Training	+	(1 | ID)	869.63	886.81	-430.82	1	48.540	< .001
Model 2	Session *d'* score	~	Training	+	(1 + Training | ID)	864.40	890.17	-426.20	2	9.237	.010
Model 3	Session *d'* score	~	Training + Age	+	(1 + Training | ID)	858.14	888.20	-422.07	1	8.259	.004
Model 4	Session *d'* score	~	Training + Age + Spatial Span BW	+	(1 + Training | ID)	855.68	890.04	-419.84	1	4.457	.035
Model 5	Session *d'* score	~	Training + Age + Spatial Span BW + Digit Span Total	+	(1 + Training | ID)	857.63	896.29	-419.82	1	0.047	.829

Note. ID: Participant ID. Training: Personal-mean-centered variable denoting the per-participant order number of a specific *N-Back* trial. Age: Grand-mean-centered variable denoting participant age as compared to the mean of all participants' ages. Spatial Span BW: Grand-mean-centered variable denoting participant WMS-III Spatial Span Backward raw score as compared to the mean of all participants' scores. Digit Span Total: A grand-mean-centered variable denoting participant WAIS-IV Digit Span total raw score as compared to the mean of all participants' scores. AIC: Akaike Information Criterion. BIC: Bayesian Information Criterion. LL: Log of likelihood. LRT Test Against Nested: Results of the Likelihood Ratio Test for the current model vs. the nested model. N of total observations for Session *d'* scores = 542. N of participants = 17. *p*-values were calculated using Satterthwaite's method.

3.7 Feasibility

The fifth research question (RQ5) was "Is participant compliance with the CognitionKit *N-Back* high-frequency testing paradigm adequate?" It was hypothesised (H5) that participant compliance would be satisfactory, in line with previous results by Cormack et al. (2022). Furthermore, it was expected that the number of missing data points would be limited to the extent as to not cause significant issues with interpretation of results.

3.7.1 Participant Compliance

Compliance was defined as the ratio of completed and scheduled *N-Back* testing sessions. Only valid and complete *N-Back* testing sessions from participants attending a specific testing block were included in the calculations. Participants were allowed to complete additional sessions of their own volition and thus compliance could exceed 100%. Comparisons of the number of scheduled and valid completed testing sessions are shown in Table 6.

Table 6. Scheduled and completed CognitionKit *N-Back* sessions.

ID	Block 1	Block 2	Block 3	Block 4	Total
N of Scheduled Sessions	204	144	156	72	576
N of Valid Completed Sessions	181	163	131	67	542
% of Scheduled Sessions Completed	88.73%	113.2%	83.97%	93.06%	94.10%
N of participants	17	12	13	6	

Note. % of Scheduled Sessions Completed corresponds to mean testing compliance.

Poor compliance was defined as a participant completing 50% or fewer of the scheduled sessions in a testing block they had agreed to complete. Poor compliance was found for only one participant, in testing blocks 2 and 3. The results supported hypothesis H5, with generally excellent compliance being demonstrated.

3.7.2 Incomplete Test Sessions and Missing Data

While participant compliance was generally high, many test sessions needed to be discarded due incomplete or missing data. 44 of the total 737 recorded CognitionKit *N-Back* test sessions were discarded, as they were mere application boots and thus contained no usable data. 131 sessions were discarded due to being non-scored practice tests completed on Mondays. 2 sessions were removed due to incorrectly

being scored tests completed on a Monday, while a non-scored practice test was scheduled. 18 sessions were removed due to incomplete data corresponding to an unfinished test. The number of data points gathered was deemed adequate, with 542 (73,5%) of the total 737 recorded test sessions across 17 participants being valid for inclusion in the analyses. Due to the robustness of the per-participant data, the number of omissions caused no significant issues with interpretation of results.

4 Discussion

The present study aimed to assess the psychometric properties and feasibility of Cambridge Cognition's CognitionKit *N-Back* (2-Back variant), a novel remotely administered n-back task designed for use in high-frequency testing. The study was conducted within a Finnish pilot sample. Contrasting the wide use of the n-back paradigm, the present literature on the psychometric properties of the task in all its variants is comparatively scarce, with many of the few studies in the literature having found middling results regarding the task's reliability and construct validity. Most of the studies have assessed the task's properties in designs including only one or two measurement sessions, however. The presently assessed CognitionKit *N-Back* task was specifically designed for use in high-frequency testing and continuous monitoring of cognitive function. As such, previous findings in the literature may thus not be directly applicable.

In the present study, *N-back* data for each participant was aggregated over each of the four week-long testing blocks respectively. Due to the averaging out of day-to-day variance, higher test-retest reliability for the task was expected as compared to the common estimates in the literature. Reliability sets the upper bounds for validity, and as such, detailed assessment of both was deemed pertinent. Age has been widely demonstrated to affect performance in both *N-Back* and other measures of WM function (e.g., Dobbs & Rule, 1989; Gajewski et al., 2018; Schmiedek, Li et al., 2009), and age-related effects were thus also accordingly analysed. As the task was specifically designed for high-frequency testing, possible training effects also needed to be assessed. Improvement through training in a task may bring additional difficulties in interpretation of results in high-frequency testing paradigms. Lastly, participant compliance was analysed, as a task can be only considered feasible for real-world use if participant engagement with the task is adequate.

4.1 Construct Validity

First, it was hypothesised (H1.1) that the assessed *N-Back* task would display adequate convergent validity by correlating strongly and significantly with other working memory measures already in wide clinical use. The strongest correlates and their possible interactions with training effects were assessed in detail. Indeed, in addition to participant age, the strongest significant predictors of *N-Back d'* in both

blocks 1 and 3 were WMS-III Spatial Span Backward scores, WMS-III Spatial Span total scores and WAIS-IV Digit Span total scores, all of which are well-validated WM tasks in common clinical use. WM measures also remained strong and significant predictors of *N-Back* performance in all multiple linear regression and repeated measures linear mixed models even when the effects of other significant covariates were statistically controlled for. Some disagreement between models existed regarding the question of which WM measure was the best predictor, however. Hypothesis H1.1 was supported, and the task was deemed to display agreeable convergent validity. These results contrast with some of the studies in the literature, as performance in n-back tasks has often been found to correlate better with measures of short-term memory (e.g., Kane et al., 2007; Roberts & Gibson, 2002; Unsworth, 2010) and executive function (e.g., Kwong See and Ryan, 1995; Miller et al., 2009) than with other measures of WM.

Hypotheses H1.2 and H1.3 concerned the discriminant validity of the *N-Back* task. It was hypothesised that correlations with common measures of executive function would be smaller in magnitude, albeit still statistically significant, and that the task would not correlate significantly with other assessed tasks. As expected in line with the literature on intercorrelations between n-back tasks and measures of executive function (e.g., Kwong See and Ryan, 1995; Miller et al., 2009), significant correlations with Stroop, TMT and WAIS-IV Coding were detected in block 1, though these findings were not able to be replicated in block 3. As hypothesised, however, the correlations between these measures of executive function and the assessed *N-Back* task were lower than the correlations between *N-Back* and measures of working memory. Surprisingly, *N-Back* performance was also found to significantly correlate with both immediate and delayed WMS-III Logical Memory. The small sample size of the present study leads to difficulties in assessing whether the non-significance of the expected correlations indicates a true absence of a relationship, or whether this discrepancy is explained by effects of training or participant dropout. The unexpected correlations between *N-Back* and WMS-III Logical Memory were judged as likely spurious, however. At present no theoretical basis for such a link could be found to exist in the literature. Statistical outliers may have extreme effects on detected correlations in samples as small as was included in the present study in block 3. While support for hypothesis H1.2 was somewhat inconclusive, hypothesis H1.3 was

supported, and the present results were deemed to suggest preliminary support for the task's discriminant validity.

4.2 Test-Retest Reliability

It was also hypothesised (H2) that the assessed *N-Back* task would display adequate test-retest reliability in a high-frequency testing paradigm when each participant's performance scores are aggregated within week-long testing blocks, averaging out day-to-day variability in performance. The hypothesis was supported, as test-retest reliability of the task was indeed found to be high. With data thusly processed, the reliability of the *N-Back* task was among the highest of all cognitive measures assessed. This result was deemed unsurprising, however, as aggregating data from multiple tests is expected to statistically raise reliability by averaging out day-to-day variation. Reliability of the task was found to be in line with the higher end of estimates found in the literature (e.g., Cormack et al., 2022; Cormack et al., 2024; Burgoyne et al., 2024).

4.3 Effects of Age

Participant age was also hypothesised (H3) to affect *N-Back* performance, as age is known to affect performance in a wide variety of working memory tasks (e.g., Dobbs & Rule, 1989; Gajewski et al., 2018) including n-back paradigms. To facilitate comparison of variables and the study of interaction effects with age, untransformed raw scores with no age standardisation were used for all cognitive measures. In addition to the significant and strong zero-order correlations between participant age and *N-Back* performance in both blocks 1 and 3, age remained a significant predictor of *N-Back* performance in all multiple linear regression and repeated measures linear mixed models even when the effects of other significant covariates were included in the models. Hypothesis H3 was thus strongly supported. These results contribute to the growing literature on aging effects in WM performance, such as n-back accuracy declining with age (Bopp & Verhaeghen, 2020). No interaction effects between age and training were found, however, contrasting previous results such as those of Salminen et al. (2016).

4.4 Training Effects

Hypothesis H4 posited that training effects would be clearly observable even in the presence of other significant predictor variables. In all assessed linear mixed models training remained a significant predictor of per-participant *N-Back* performance. I.e., performance within-participants became better over time with an increased number of repetitions of the task. Hypothesis H4 was thus supported, and the findings of the present study contribute to plentiful examples of effects of training on *N-Back* performance in the literature (e.g., Li et al., 2008)

4.5 Feasibility

Lastly, the present study assessed the real-world feasibility of high-frequency *N-Back* testing by assessing participant testing compliance and patterns of missing data. Supporting hypothesis H5, participant compliance was found to be generally very high, with poor compliance demonstrated in only one participant. Though the personal datasets of most participants included some missing data, the omissions were generally quite minor and were of no consequence to the analyses performed due to the robustness of the whole dataset.

4.6 Other Findings

Interestingly, the correlation between *N-Back* accuracy and WAIS-IV Digit Span Backward was significant at testing block 3 but not block 1, while the correlation between *N-Back* accuracy and WAIS-IV Digit Span Sequencing was significant at block 1 but not block 3. Both tasks are well-validated measures of WM in common clinical use, and as such would be expected to correlate significantly with performance in other WM tasks. By extension, if *N-Back* is to be considered a WM task in the same vein as these tasks, significant correlations would be expected. It is important to note, however, that small sample sizes notably raise the bar for magnitude of effects to achieve statistical significance. Correlation coefficients exceeding $r = .30$ are traditionally considered medium in their magnitude (Hemphill, 2003; Cohen, 1988, as cited in Hemphill, 2003), yet failed to reach significance in the present, small sample. Additionally, WAIS-IV Digit Span Total Raw score included performance in both of these tasks. This combined score remained a significant predictor of *N-Back* performance in both of the cross-sectional analyses at blocks 1

and 3, however. In sum, the results on the statistical significance and magnitude of the relevant correlations lent support for the hypothesis H1.1, with *N-Back* displaying adequate convergent validity with other relevant WM measures despite these unexpected inconsistencies.

4.7 Limitations and Future Directions

As mentioned, the size of the present sample was quite small owing to the pilot nature of the study. Due to this, some of the detected significant correlations and the absences of expected ones could have been spurious. This was judged as a likely scenario regarding the unexpected correlation between *N-Back* accuracy and WMS-III Logical Memory, as no theoretical basis for such a link could be found in the literature. A larger sample would also enable the use of factor analysis methods to further assess the construct validity of *N-Back*.

Age range (28–64), educational level, language and gender distributions of participants studied were also somewhat restricted, with most participants being highly educated, middle-aged Finnish-speaking women. The present findings may thus not be completely generalisable to the population at large. The importance of diverse samples in studies on psychometric measures has been noted prior by e.g., Redick and Lindsey (2013). Diversity of educational background is especially important for generalisability, as restriction of range associated with educational level may affect results of correlational analyses of cognitive measures. Correlations between psychometric measures have been noted to differ in different groups of varying academic and cognitive ability (e.g., Blum & Holling, 2017), a series of findings originating in Spearman's (1927) Law of Diminishing Returns.

Tiredness and fatigue have also been shown to affect cognitive performance. The present study was restricted to assessing these possible within-participant fatigue related effects using the somewhat crude measure of time of day at time of test. Regrettably, data for participant's specific sleep schedules were not available to be analysed in the present study. Detailed assessment of wakefulness-related differences in *N-Back* performance between e.g., night shift workers and day shift workers was thus not possible. No significant differences between morning and evening scores were detected in the present study, however. This surprising finding on assumed tiredness not affecting *d'* scores would benefit from further research attention.

4.8 Conclusions

Most of the presented hypotheses received considerable support from the data collected in this Finnish pilot sample. Though the size of the present sample was quite small, the per-participant data gathered was remarkably robust and comprehensive both in breadth and depth. Although replications in larger and more diverse samples are certainly needed, the present results indicate that the assessed *N-Back* task displays adequate test-retest reliability and construct validity. Age was also found to affect performance in the task as expected for a measure of WM. Training affected within-participant performance, with performance improving over repetitions of the task. Training effects need to be taken into consideration when interpreting longitudinal results on per-participant performance. In addition to the satisfactory psychometric qualities exhibited by the task, participant testing compliance was found to be excellent. Use of the task in high-frequency testing as designed may be deemed feasible according to the results of the present pilot study.

References

Ahn, Y. D., Yi, D., Joung, H., Seo, E. H., Lee, Y. H., Byun, M. S., Lee, J.H., Jeon, S.Y., Lee, J.Y., Sohn, B.K., Lee, D.Y., & KBASE Research Group. (2019). Normative data for the logical memory subtest of the Wechsler Memory Scale-IV in middle-aged and elderly Korean people. *Psychiatry investigation*, *16*(11), 793.

Amieva, H., Lafont, S., Auriacombe, S., Rainville, C., Orgogozo, J. M., Dartigues, J. F., & Fabrigoule, C. (1998). Analysis of error types in the Trail Making Test evidences an inhibitory deficit in dementia of the Alzheimer type. *Journal of clinical and experimental neuropsychology*, *20*(2), 280-285.

Arbuthnott, K., & Frank, J. (2000). Trail making test, part B as a measure of executive control: validation using a set-switching paradigm. *Journal of clinical and experimental neuropsychology*, *22*(4), 518-528.

Army Individual Test Battery. (1944). *Manual of Directions and Scoring*. War Department, Adjutant General's Office.

Baddeley, A. (1992). Working memory: The interface between memory and cognition. *Journal of cognitive neuroscience, 4*(3), 281-288.

Bopp, K. L., & Verhaeghen, P. (2020). Aging and n-back performance: A meta-analysis. *The Journals of Gerontology: Series B*, *75*(2), 229-240.

Burgoyne, A. P., Frank, D. J., & Macnamara, B. N. (2024). Which "working memory" are we talking about? Complex span tasks versus N-back. *Psychonomic Bulletin & Review*, 1-15.

Cambridge Cognition. (2023). *Cognition Kit N-Back (NBX).* https://cambridgecognition.com/cognitionkit-n-back-nbx/

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81.

Ciesielski, K. T., Lesnik, P. G., Savoy, R. L., Grant, E. P., & Ahlfors, S. P. (2006). Developmental neural networks in children performing a Categorical N-Back Task. *Neuroimage*, *33*(3), 980-990.

Cohen, J. (1998). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale.

Colliver, J. A., Conlee, M. J., & Verhulst, S. J. (2012). From test validity to construct validity... and back?. *Medical education*, *46*(4), 366-371.

Colom, R., Abad, F. J., Quiroga, M. Á., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs, but why?. *Intelligence*, *36*(6), 584-606.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic bulletin & review*, *12*, 769-786.

Cormack, F., McCue, M., Skirrow, C., Cashdollar, N., Taptiklis, N., van Schaik, T., Fehnert, B., King, J., Chrones, L., Sarkey, S., Kroll, J., & Barnett, J. (2024). Characterizing longitudinal patterns in cognition, mood, and activity in depression with 6-week high-frequency wearable assessment: Observational study. *JMIR Mental Health*, *11*(1), e46895.

Cormack, F., Ticcinelli, V., Taptiklis, N., Kudelka, J., Emmert, K., Maetzler, W., Reilmann, R., Latzman, R. D., Ng, W. F., McRae, V., Davies, K., van der Woude, J., Fierrez, J., Ahmaniemi, T., & Chatterjee, M. (2022). App-based cognitive assessment and monitoring: a feasibility study in patients with immune-mediated inflammatory and neurodegenerative disorders. *Neurosci Appl*, *1*, 100826.

Cowan, N. (2012). *Working memory capacity*. Psychology press.

Dobbs, A. R., & Rule, B. G. (1989). Adult age differences in working memory. *Psychology and aging*, *4*(4), 500.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of experimental psychology: General*, *128*(3), 309.

Espeland, M. A., Katula, J. A., Rushing, J., Kramer, A. F., Jennings, J. M., Sink, K. M., ... & LIFE Study Group. (2013). Performance of a computer-based assessment of cognitive function measures in two cohorts of seniors. *International journal of geriatric psychiatry*, *28*(12), 1239-1250.

Fraga, F. J., Mamani, G. Q., Johns, E., Tavares, G., Falk, T. H., & Phillips, N. A. (2018). Early diagnosis of mild cognitive impairment and Alzheimer's with event-related potentials and event-related desynchronization in N-back working memory tasks. *Computer methods and programs in biomedicine*, *164*, 1-13.

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological science*, *17*(2), 172-179.

Frost, A., Moussaoui, S., Kaur, J., Aziz, S., Fukuda, K., & Niemeier, M. (2021). Is the n-back task a measure of unstructured working memory capacity? Towards understanding its connection to other working memory tasks. *Acta Psychologica*, *219*, 103398.

Gajewski, P. D., Hanisch, E., Falkenstein, M., Thönes, S., & Wascher, E. (2018). What does the n-back task measure as we get older? Relations between working-memory measures and other cognitive functions across the lifespan. *Frontiers in psychology*, *9*, 2208.

Gevins, A., & Smith, M. E. (2000). Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. *Cerebral cortex*, *10*(9), 829-839.

Gluud, L. L., Jeyaraj, R., & Morgan, M. Y. (2019). Outcomes in clinical trials evaluating interventions for the prevention and treatment of hepatic encephalopathy. *Journal of Clinical and Experimental Hepatology*, *9*(3), 354-361.

Goldberg, E., & Bougakov, D. (2005). Neuropsychologic assessment of frontal lobe dysfunction. *Psychiatric Clinics*, *28*(3), 567-580.

Goode, W. J., & Hatt, P. K. (1952). *Methods in social research*. McGraw Hill.

Groth-Marnat, G., & Wright, A. J. (2016). *Handbook of psychological assessment* (6th ed.). John Wiley & Sons.

Haatveit, B. C., Sundet, K., Hugdahl, K., Ueland, T., Melle, I., & Andreassen, O. A. (2010). The validity of d prime as a working memory index: results from the "Bergen n-back task". *Journal of clinical and experimental neuropsychology*, *32*(8), 871-880.

Hammersley, M. (1987). Some notes on the terms 'validity' and 'reliability'. *British educational research journal*, *13*(1), 73-82.

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients.

Henik, A., & Salo, R. (2004). Schizophrenia and the Stroop effect. *Behavioral and cognitive neuroscience reviews*, *3*(1), 42-59.

Hockey, A., & Geffen, G. (2004). The concurrent validity and test–retest reliability of a visuospatial working memory task. *Intelligence*, *32*(6), 591-605.

Jääskeläinen, E., Haapea, M., Juola, T., Rannikko, I., Rautio, N., Lehto, S., Ylitalo, J., Koponen, H., Isohanni, M., Murray, G., Jones, P., & APSY Oulu Group. (2021). Benefits and risks of off label use of antipsychotics in insomnia and anxiety–APSY Oulu project. *Nordic Journal of Psychiatry*, *75*(sup1), S16-S16.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, *18*(4), 394-412.

Jonides, J., Schumacher, E. H., Smith, E. E., Lauber, E. J., Awh, E., Minoshima, S., & Koeppe, R. A. (1997). Verbal working memory load affects regional brain activation as measured by PET. *Journal of cognitive neuroscience*, *9*(4), 462-475.

Joung, H., Yi, D., Byun, M. S., Lee, J. H., Lee, Y., Ahn, H., & Lee, D. Y. (2021). Functional neural correlates of the WAIS-IV Block Design Test in older adult with mild cognitive impairment and Alzheimer's Disease. *Neuroscience*, *463*, 197-203.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic bulletin & review*, *9*(4), 637-671.

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the N-back task: a question of construct validity. *Journal of Experimental psychology: learning, memory, and cognition*, *33*(3), 615.

Kaplan, E., Fein, D., Kramer, J., Delis, D., & Morris, R. (1999). *Manual of the Wechsler Intelligence Scale for Children Third Edition (WISC-III) as a process instrument*. The Psychological Corporation.

Kessels, R. P., Overbeek, A., & Bouman, Z. (2015). Assessment of verbal and visuospatial working memory in mild cognitive impairment and Alzheimer's dementia. *Dementia & Neuropsychologia*, *9*, 301-305.

Kwong See, S. T., & Ryan, E. B. (1995). Cognitive mediation of adult age differences in language performance. *Psychology and aging*, *10*(3), 458.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?!. *Intelligence*, *14*(4), 389-433.

Li, S. C., Schmiedek, F., Huxhold, O., Röcke, C., Smith, J., & Lindenberger, U. (2008). Working memory plasticity in old age: practice gain, transfer, and maintenance. *Psychology and aging*, *23*(4), 731.

Manglam, M. K., Ram, D., Praharaj, S. K., & Sarkhel, S. (2010). Working memory in schizophrenia. *Age*, *16*(6.34).

McMillan, K. M., Laird, A. R., Witt, S. T., & Meyerand, M. E. (2007). Self-paced working memory: Validation of verbal variations of the n-back paradigm. *Brain research*, *1139*, 133-142.

Mendez, M. F. (2021). *The Mental Status Examination Handbook E-Book*. Elsevier Health Sciences.

Miller, K. M., Price, C. C., Okun, M. S., Montijo, H., & Bowers, D. (2009). Is the n-back task a valid neuropsychological measure for assessing working memory?. *Archives of Clinical Neuropsychology*, *24*(7), 711-717.

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (1999). *Handbook of normative data for neuropsychological assessment*. Oxford University Press.

Miyake, A., & Shah, P. (1999). *Models of working memory: Mechanisms of active maintenance and executive control*, 442-481. Cambridge University Press.

Nybo, T., Akila., R., & Kärki., P. (2000). *Ikäryhmäviiteaineisto Stroop- ja Raven Advanced Progressive Matrices -testeille*. Aivotyölaboratorio, Työterveyslaitos.

Oberauer, K. (2005). Binding and inhibition in working memory: individual and age differences in short-term recognition. *Journal of experimental psychology: General*, *134*(3), 368.

Olchik, M. R., Ghisi, M., Freiry, A. M., Ayres, A., Schuh, A. F. S., Rieder, C. R. D. M., & Teixeira, A. R. (2017). Comparison trail making test between individuals with Parkinson's disease and health controls: suggestions of cutoff point. *Psychology & Neuroscience, 10*(1), 77.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping*, *25*(1), 46-59.

Perry, W., Heaton, R. K., Potterat, E., Roebuck, T., Minassian, A., & Braff, D. L. (2001). Working memory in schizophrenia: transient "online" storage versus executive functioning. *Schizophrenia bulletin, 27*(1), 157-176.

Poutiainen, E., Kalska, H., Laasonen, M., Närhi, V., & Räsänen, P. (2010). *Trail making-testi: Käsikirja*. Psykologien Kustannus Oy.

Redick, T. S., & Lindsey, D. R. (2013). Complex span and n-back measures of working memory: A meta-analysis. *Psychonomic bulletin & review*, *20*, 1102-1113.

Reinvall, O., & Poutiainen, E. (2008). Tapping -tehtävän suomalainen viitearvoaineisto. Verkkojulkaisu. Helsinki: Suomen Neuropsykologinen Yhdistys ry.

Roberts, R., & Gibson, E. (2002). Individual differences in sentence memory. *Journal of psycholinguistic research*, *31*, 573-598.

Ryan, J. J., Sumerall, S. W., Seeley, J. S., Umfleet, L. G., Kreiner, D., Brown, K. I., & Ott, S. D. (2015). WAIS–IV coding performance of young adults: transcription patterns and incidental learning procedures. *North American Journal of Psychology*, *17*(1), 197-212.

Salminen, T., Frensch, P., Strobach, T., & Schubert, T. (2016). Age-specific differences of dual n-back training. *Aging, Neuropsychology, and Cognition*, *23*(1), 18-39.

Scarpina, F., & Tagini, S. (2017). The Stroop color and word test. *Frontiers in psychology*, *8*, 557.

Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: the gap is not that deep. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 1089.

Schmiedek, F., Li, S. C., & Lindenberger, U. (2009). Interference and facilitation in spatial working memory: age-associated differences in lure effects in the n-back paradigm. *Psychology and aging*, *24*(1), 203.

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2014). A task is a task is a task: Putting complex span, n-back, and other working memory indicators in psychometric context. *Frontiers in psychology*, *5*, 1475.

Shamosh, N. A., DeYoung, C. G., Green, A. E., Reis, D. L., Johnson, M. R., Conway, A. R., Engle, R.W., Braver, T.S., & Gray, J.R., (2008). Individual differences in delay discounting: relation to intelligence, working memory, and anterior prefrontal cortex. *Psychological science*, *19*(9), 904-911.

Shelton, J. T., Elliott, E. M., Hill, B. D., Calamia, M. R., & Gouvier, W. D. (2009). A comparison of laboratory and clinical working memory tests and their prediction of fluid intelligence. *Intelligence*, *37*(3), 283-293.

Shelton, J. T., Metzger, R. L., & Elliott, E. M. (2007). A group-administered lag task as a measure of working memory. *Behavior Research Methods*, *39*, 482-493.

Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A multi-level meta-analysis of n-back training studies. *Psychonomic bulletin & review*, *24*, 1077-1096.

Spearman, C. (1927/1961). *The abilities of man*. Macmillan.

Stenberg, J., Laari, S., Uimonen, J., Pihlaja, R., & Poutiainen, E. (2016). Työikäisten kognitiivinen suoriutuminen–viitearvotietoja AINO-tutkimuksesta. *Psykologia*, *51*(6), 400-421.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*(6), 643.

The Psychological Corporation. (1997). *The WAIS-III–WMS-III Technical Manual*. The Psychological Corporation.

The Psychological Corporation. (2002). *Updated WAIS-III–WMS-III Technical Manual*. The Psychological Corporation.

Tombaugh, T. N. (2004). Trail Making Test A and B: normative data stratified by age and education. *Archives of clinical neuropsychology*, *19*(2), 203-214.

Tulsky, D. S. (2004). A new look at the WMS-III: new research to guide clinical practice. *Journal of clinical and experimental neuropsychology*, *26*(4), 453-458.

Tyni, T. (2022). *WMS-III:n osatestien ja indeksien suoriutumiserojen reliabiliteetit ja luottamusvälit* (Master's thesis, Itä-Suomen yliopisto).

Unsworth, N. (2010). On the division of working memory and long-term memory and their relation to intelligence: A latent variable approach. *Acta psychologica*, *134*(1), 16-28.

Uttl, B., Graf, P., & Richter, L. K. (2002). Verbal paired associates tests limits on validity and reliability. *Archives of Clinical Neuropsychology*, *17*(6), 567-581.

Watter, S., Geffen, G. M., & Geffen, L. B. (2001). The n-back as a dual-task: P300 morphology under divided attention. *Psychophysiology*, *38*(6), 998-1003.

Wechsler, D. (1997). *The Wechsler Memory Scale-III*. The Psychological Corporation.

Wechsler, D. (2008a). *Wechsler Adult Intelligence Scale* (4th ed.). Pearson Assessment.

Wechsler, D. (2008b). *WMS-III käsikirja*. Psykologien Kustannus Oy.

Wechsler, D. (2012). *WAIS-IV käsikirja (esitys- ja pisteytyskäsikirja)*. Psykologien Kustannus Oy.

Weiss, L. G., Saklofske, D. H., Holdnack, J. A., & Prifitera, A. (2016). *WISC-V assessment and interpretation: Scientist-practitioner perspectives*. Elsevier Academic Press.

Wilde, N. J., Strauss, E., & Tulsky, D. S. (2004). Memory span on the Wechsler scales. *Journal of clinical and experimental neuropsychology*, *26*(4), 539-549.

Wilde, N., & Strauss, E. (2002). Functional equivalence of WAIS-III/WMS-III digit and spatial span under forward and backward recall conditions. *The Clinical Neuropsychologist*, *16*(3), 322-330.

Yang, Y., Conners, F. A., & Merrill, E. C. (2014). Visuo-spatial ability in individuals with Down syndrome: Is it really a strength?. *Research in developmental disabilities*, *35*(7), 1473-1500.