Published in final edited form as:

J Clin Exp Neuropsychol. 2014 October ; 36(8): 875–886. doi:10.1080/13803395.2014.953039.

Clinical Utility of the N-back Task in Functional Neuroimaging Studies of Working Memory

Lisa M. Jacola,

Department of Psychology, St. Jude Children's Research Hospital, 262 Danny Thomas Way, Memphis, Tennessee, 38105

Victoria W. Willard,

Department of Psychology, St. Jude Children's Research Hospital, 262 Danny Thomas Way, Memphis, Tennessee, 38105

Jason M. Ashford,

Department of Psychology, St. Jude Children's Research Hospital, 262 Danny Thomas Way, Memphis, Tennessee, 38105

Robert J. Ogg,

Division of Translational Imaging, St. Jude Children's Research Hospital, 262 Danny Thomas Way, Memphis, Tennessee, 38105

Matthew A. Scoggins,

Division of Translational Imaging, St. Jude Children's Research Hospital, 262 Danny Thomas Way, Memphis, Tennessee, 38105

Melissa M. Jones,

Division of Translational Imaging, St. Jude Children's Research Hospital, 262 Danny Thomas Way, Memphis, Tennessee, 38105

Shengjie Wu, and

Department of Biostatistics, St. Jude Children's Research Hospital, 262 Danny Thomas Way, Memphis, Tennessee, 38105

Heather M. Conklin

Department of Psychology, St. Jude Children's Research Hospital, 262 Danny Thomas Way, Memphis, Tennessee, 38105

Abstract

Introduction—N-back tasks are commonly used in functional neuroimaging studies to identify the neural mechanisms supporting working memory (WM). Despite widespread use, the clinical utility of these tasks is not well specified. This study compared N-back performance during functional magnetic resonance imaging (fMRI) with task data acquired outside of the scanner as a

Corresponding Author: Heather M. Conklin, PhD, Department of Psychology, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Mail Stop # 740, Memphis, Tennessee, 38105, Telephone: 901-595-3585, Fax: 901-595-4701, heather.conklin@stjude.org.

Disclosure Statement: The authors have no financial interest or benefit in the direct application of this research.

measure of reliability across environment. N-back task validity was examined in relation to performance and rater-based measures used clinically to assess working memory.

Method—Forty-three healthy adults completed Verbal and Object N-back tasks during fMRI scanning and outside the scanner. Task difficulty was varied parametrically (0, 1, and 2-back conditions). Order of N-back task completion was stratified by modality (Verbal/Object) and environment. Participants completed the Digit Span [DS] and provided self-ratings using the Behavior Rating Inventory of Executive Function [BRIEF-WM]).

Results—Mean Verbal and Object N-back accuracy was above 95% across load conditions; task difficulty was effectively manipulated across load conditions. Performance accuracy did not significantly differ by environment. N-back reaction time was slower during fMRI ($F = 6.52$, $p = .01$, $\eta_p^2 = .13$); participants were faster when initially completing tasks outside the scanner ($\eta_p^2 = .10$ –.15). Verbal 2-back accuracy was significantly related to DS performance ($r = .36$, $p = .02$). N-back performance was not related to BRIEF-WM.

Conclusions—Our results provide evidence for reliability of N-back accuracy during fMRI scanning; however, reliability of reaction time data is affected by order of task presentation. Data regarding construct validity are inconsistent and emphasize the need to consider clinical utility of behavioral measures in the design and interpretation of functional neuroimaging studies.

Keywords

Digit Span; N-back; BRIEF; working memory; validity; reliability

Introduction

Working memory is a core neuropsychological function responsible for facilitating online storage and manipulation of information used to guide cognition and behavior (Baddeley, 1986). Studies have consistently identified working memory ability as an important component of complex academic skills, including language comprehension, mathematical computation, reading comprehension, and written expression (Ayr, Yeates, & Enrile, 2005; de Jonge & de Jong, 1996; Hanten, Levin, & Song, 1999; Swanson, 1999). Developmental improvements in working memory have been shown to account for a significant proportion of age-related improvements in IQ (Fry & Hale, 1996).

In clinical settings, working memory is typically assessed using recall-based tasks that involve rehearsal and manipulation of information over a brief period of time, such as the Digit Span subtest from the Wechsler Adult Intelligence Scales, 3rd Edition (WAIS-III; Wechsler, 1997). The Digit Span subtest involves Digit Span Forward, in which individuals are required to repeat verbatim number sequences read by the examiner, and Digit Span Backward, which requires participants to mentally rehearse and manipulate verbally presented sequences of numbers prior to repeating them backward. Forward span tasks are considered measures of auditory attention and immediate recall, while backward span tasks more specifically tap working memory skills (Lezak, 1995).

Rater-based measures, including informant and self-report questionnaires, are also used clinically in order to gain information regarding problems with neurocognitive function. The

Behavior Rating Inventory of Executive Function, Adult (BRIEF-A) is a rater-based questionnaire that is used to assess whether executive functions, including working memory, are problematic in the context of everyday functioning (Roth, Isquith, & Gioia, 2005). The association of performance and rater-based measures of working memory has been inconsistent in the existing research literature. Some studies find that subjective ratings are more sensitive to executive dysfunction than traditional, performance-based measures (Dawson, Shear, & Strakowski, 2012; Miyake, Emerson, & Friedman, 2000; Miyake, Friedman, et al., 2000; Nigg et al., 2005). In practice, clinicians often make use of both performance and rater-based measures, concluding that both types of instruments yield complementary information that accounts for different proportions of variability.

In contrast to clinical assessment, where the goal is often to provide an individualized conceptualization of a patient's neurobehavioral profile and provide recommendations to promote functioning across multiple domains, functional neuroimaging studies of cognition aim to identify patterns of neural activation that support cognitive processes. As such, researchers use behavioral measures that reliably elicit patterns of neural activation and can easily be correlated with the time course of cognitive processing. Computerized measures such as the N-back are well suited for this purpose, as they allow for parametric variation of task difficulty, multimodal stimulus presentation, and precise measurement of performance characteristics. N-back tasks require participants to respond to a presented stimulus only when it is the same as one presented on a trial at a predetermined number (N) prior to the current trial (see Figure 1); as such, the N-back requires constant online maintenance, updating, discrimination, and recognition, in order to accurately respond to target stimuli that recur at the specified interval.

N-back tasks have become prototypical measures in functional neuroimaging studies of working memory. Studies consistently find that N-back performance is associated with activation in prefrontal and parietal cortical regions widely recognized as the primary neural substrates that underlie working memory processes (Garlinghouse, Roth, Isquith, Flashman, & Saykin, 2010; Goldman-Rakic, 1995; Grön, 1998; Mahone, Martin, Kates, Hay, & Horská, 2009; Owen, McMillan, Laird, & Bullmore, 2005; Petrides, 1995; Smith & Jonides, 1998). Patterns of neural activation associated with N-back performance have been shown to vary with the type of information held in working memory (e.g., verbal or spatial), as well as task difficulty (0, 1, 2-back; see Rottschy et al., 2012; Smith & Jonides, 1998 for reviews).

Despite being frequently used in clinical research, there have been limited psychometric studies of the N-back task. Studies exploring the reliability of N-back performance outside of the scanner have yielded variable findings. As an example, Jaeggi and colleagues (2010) examined the split-half reliability of performance on auditory and visual N-back tasks in three different samples. Reliability coefficients ranged from .09 to .08 in the two-back condition and .39 to .60 in the 3-back conditions. In contrast, a number of studies included in a recent meta-analysis of N-back performance reported reliability coefficients greater than .70 (see Redick & Lindsey, 2013).

Reliability concerns are particularly relevant in functional neuroimaging research, where researchers must explain and control for multiple sources of variability in data. For example,

data collected using functional magnetic resonance imaging (fMRI) are influenced by factors related to the scanner (e.g., magnet strength, signal-to-noise properties, and electronic signal drift), participants (e.g., characteristics of a cohort, physiological signals, movement, and emotional state of individual participants), and tasks (e.g., reliability, training effects, and design characteristics), all of which are potential sources of variability (Bennett & Miller, 2010; Gaillard, Grandin, & Xu, 2001; Handwerker, Ollinger, & D'Esposito, 2004; Lueken, Muehlhan, Wittchen, & Kirschbaum, 2012; Plichta et al., 2012). The few studies that have directly examined test-retest reliability of N-back performance during fMRI have raised concerns about the effect of practice on task performance and subsequently the strength of neural activation (Bennett & Miller, 2010; Plichta et al., 2012). An increased understanding of the sources of between and within-group variability will add information that can improve the design of multi-center studies, as well as fMRI used for clinical purposes, such as presurgical planning and the assessment of therapeutic efficacy (Barch & Mathalon, 2011; Glover et al., 2012; Szaflarski et al., 2008; Zou et al., 2012).

The widespread use of N-back tasks as a measure of working memory is also concerning given inconsistent findings regarding construct validity. Studies have generally found that N-back performance is either weakly correlated with or unrelated to commonly used clinical measures of working memory, including Digit Span Backward, reading span, and math-based span tasks (Kane et al., 2007; Miller, Price, Okun, Montijo, & Bowers, 2009; Oberauer et al., 2005; Roberts & Gibson, 2002). Studies examining the relationship between N-back performance and global intelligence have identified moderate to weak associations between N-back accuracy and measures of intelligence (Friedman, Miyake, Corley, Young, & Hewitt, 2006; Friedman et al., 2008; Shelton, Elliott, Hill, Calamia, & Gouvier, 2009; Van Leeuwen, Van den Berg, Hoekstra, & Boomsma, 2007; Waiter et al., 2009). Reaction time has been shown to predict individual differences in measured intelligence, such that those with higher measured intelligence responded more quickly during N-back trials (Gevins & Smith, 2000; Hockey & Geffen, 2004). Reaction time has also been shown to successfully differentiate between healthy and clinical adult cohorts, including groups of individuals with prefrontal dysfunction, such as schizophrenia and Parkinson's disease (Miller et al., 2009; Perlstein, Carter, Noll, & Cohen, 2001).

Findings of inconsistent validity are particularly notable given the increased use of computerized measures as clinically acceptable measures of neuropsychological functioning. While computerized measures are advantageous in terms of time and ease of administration, questions remain regarding their clinical utility in relation to traditionally accepted measures of neuropsychological constructs. It is incumbent on clinicians to ensure that these instruments measure similar constructs prior to using them to answer clinically driven questions regarding current level of functioning or treatment outcomes.

To our knowledge, no study has directly compared task performance during fMRI scanning to performance outside of the scanner. Thus, the primary aim of this study was to explore the clinical utility of the N-back as a measure of working memory during fMRI scanning in a group of healthy adults. We hypothesized that performance reliability would not significantly differ across environment. Validity was examined in relation to measures of intelligence and working memory commonly used in clinical settings, including the

Wechsler Abbreviated Scales of Intelligence (WASI; Wechsler, 1999), the Digit Span subtest from the WAIS-III (Wechsler, 1997) and ratings of working memory problems from the self-report form of the BRIEF-A (Roth et al., 2005). We hypothesized that N-back performance would significantly but modestly correlate with Digit Span Backward, the Working Memory clinical scale of the BRIEF-A, and an abbreviated measure of intelligence (WASI Abbreviated IQ).

Methods

Participants

Data were collected as part of a larger fMRI study that sought to establish typical patterns of neural activation during working memory tasks in healthy young adults for eventual comparison with clinical populations. Forty-five adult participants were recruited from the community to complete neuropsychological assessment and fMRI scans. Enrollment was stratified by gender. All participants were right-handed and primary English speakers. Individuals with a reported history of Attention Deficit Hyperactivity Disorder (ADHD), Learning Disorder, or significant impairment in intellectual function (history of pull-out services or education within a self-contained classroom) were not considered for participation. Additional exclusion criteria included a reported history of central nervous system injury or disease, mental illness associated with brain changes (e.g., schizophrenia, obsessive-compulsive disorder), use of stimulant or psychotropic medication within two weeks of study enrollment, prior treatment for alcohol or drug abuse, and sensory or motor impairment that would impact the validity of testing. Female participants were excluded if they were pregnant at the time of enrollment. The study was approved by the Institutional Review Board at our institution. Participants were paid 40 dollars upon completion of all procedures to compensate for time and any travel expenses.

Demographic Characteristics

Demographic characteristics are detailed in Table 1. Self-report data (responses on the demographic form and BRIEF-A) were collected for 45 participants who were between 18 and 30 years of age at the time of participation. The group was balanced with respect to gender ($\chi^2 = 0.12$, $p = .73$, $v = .02$). The majority of the sample self-identified as Caucasian and non-Hispanic. fMRI data were not collected for one participant due to equipment issues. An additional participant did not complete the Digit Span, WASI, and N-back task outside the scanner. Thus, group data differed by one participant for measures collected during fMRI and outside the scanner, resulting in data from 43 participants being analyzed for reliability comparisons.

Table 2 includes descriptive characteristics regarding performance on clinical measures of working memory and intelligence. Global intellectual ability ranged from average to superior when compared to same aged peers (WASI ABIQ range = 101 to 134). Mean performance was not significantly different on the Vocabulary and Matrix Reasoning subtests ($t(43) = -1.05$, $p = .30$, $d = -0.22$). Digit Span performance ranged from borderline impaired to well above age expectations (Digit Span Total Scaled Score range = 6 to 18). There were no clinically significant mean elevations on BRIEF-A Composite or

Clinical scales, including the working memory scale (BRIEF-WM). To more closely examine the rate of self-reported executive functioning difficulties, participants were categorized as elevated ($T > 60$) or not elevated ($T \leq 59$) on the BRIEF-WM scale, using a cutoff of 1 *SD*. Results of categorical comparisons revealed no difference between the proportion of sample participants at risk for significant working memory problems when compared to normative expectations (11% and 16%, respectively; $\chi^2 = 1.13$, $p = .29$, $v = 0.02$).

Procedure

Written informed consent was obtained from all participants at the beginning of the study visit. After being consented, participants completed a self-report questionnaire to gain information regarding general demographics (e.g., age, gender, ethnicity, and mental health diagnoses) and family demographic characteristics (e.g., parental education and occupation), in order to derive an index of socioeconomic status. Participants then completed an assessment using clinical measures of working memory and intelligence. There was no prescribed order of completion for these tasks.

All participants completed anatomical scans lasting 9 minutes prior to fMRI tasks. Participants completed a total of four tasks in a session. Two of these are not discussed in the manuscript; the length of these two tasks was dependent on participant performance and lasted between 4 and 10 minutes. N-back task order was randomized across environment (initial exposure during fMRI or outside the scanner) and task modality (initial task Verbal or Object N-back). Order of task presentation was randomized in an attempt to balance concerns regarding the differential impact of fatigue, etc.. Specifically, 50% of participants completed either the Verbal or Object N-back immediately after the anatomical scanning and 50% of the participants completed anatomical scanning plus 4 to 10 minutes of additional tasks prior to completing the N-back tasks. Tasks inside and outside the scanner were completed on the same day.

Regardless of task order, all participants first viewed a Power Point presentation that included pictures of the MRI scanner and a brief description of study procedure prior to completing the N-back tasks. The instructions provided in this training presentation described the task without emphasizing speed or accuracy (e.g., Verbal N-back: "The letters come in blocks of 16 each. Before each block you will see an instruction telling you the interval for the next block. Click the mouse when you see a letter appear or reappear at that specified interval."). Participants also completed a one-block trial of the Verbal and Object N-back tasks, which provided an opportunity to practice with the pneumatic squeeze ball response mechanism used during fMRI data collection.

Measures

Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999)—The Vocabulary and Matrix Reasoning subtests from the WASI were administered in order to obtain an abbreviated IQ (ABIQ). ABIQ scores are age standardized with a mean (*M*) of 100 and a standard deviation (*SD*) of 15. Scores on the Vocabulary and Matrix Reasoning subtests are age standardized T-scores ($M = 50$, $SD = 10$). The ABIQ correlates highly with

the Full Scale IQ score from the WAIS-III ($r = .87$). Internal consistency reliability and test retest reliability for the ABIQ are high ($r = .96$ and $r = .88$, respectively).

Behavior Rating Inventory of Executive Function, Adult Version (BRIEF-A; Roth et al., 2005)—The BRIEF-A is a self-report questionnaire designed to assess behavioral manifestations of executive function in daily life. The BRIEF-A consists of 75 items from which nine clinical scales (Inhibit, Shift, Emotional Control, Initiate, Working Memory, Plan/Organize, Organization of Materials, Self-Monitor and Task Monitor), two index scores (Metacognition and Behavioral Regulation) and one composite score (Global Executive Composite) are derived. Scores are age and gender standardized ($M = 50$, $SD = 10$), with T scores greater than 65 indicative of clinically significant concerns. Internal consistency is moderate to high for all clinical scales ($\alpha s = .80 - .94$), including the working memory scale ($\alpha = .80$). Test-retest correlations were high across the clinical scales ($r s = .82 - .93$) including the working memory scale individually ($r = .92$) (Roth et al., 2005). The BRIEF-A demonstrates significant correlations in the expected directions with the Frontal Systems Behavior Scale, Dysexecutive Questionnaire and Cognitive Failures Questionnaire (Roth et al., 2005). Validity has further been demonstrated in individuals diagnosed with ADHD, as well as those with neurological impairment (i.e. epilepsy and traumatic brain injury; Roth et al., 2005; Roth, Lance, Isquith, Fischer, & Giancola, 2013).

Digit Span—Participants were administered the Digit Span subtest from the WAIS-III (Wechsler, 1997) as a measure of working memory. Z-scores were computed separately for the raw performance scores on the longest digit span forward (LDSF) and longest digit span backward (LDSB) using M and SD data provided in the WAIS-III manual (Wechsler, 1997). Internal consistency reliability and test-retest reliability for this subtest are high ($r = .87$, $r = .83$, respectively).

Verbal and Object N-back tasks—Participants completed the Verbal and Object N-back tasks during fMRI and outside of the scanner (Figure 1). For the Verbal tasks, participants viewed a continuous stream of single random stimuli from a set of phonologically distinct letters and responded when the currently presented letter was identical to the letter presented at the specified interval (1 or 2-back). The Object N-back required participants to monitor a continuous stream of single objects and select matches in the same manner as the Verbal tasks. A control condition (0-back) was used during which the same continuous stream of single letters (Verbal) or objects (Object) were presented.

Each N-back task paradigm (Verbal/Object) consisted of three blocks of stimuli. Each block contained 4 targets and 12 distractors for 0, 1, and 2-back portions, respectively. This resulted in a total of 12 targets and 36 distractors for task load across one 4-minute paradigm. Participants were informed of a target interval change via visually presented instructions: “Press the button when you see X” for 0-back; “Same as the one before?” for 1-back; “Every other one?” for 2-back. Stimuli were presented on a computer monitor for 0.5 seconds, with an inter-stimulus interval of 1.5 seconds. Each participant completed four N-back paradigms: Verbal and Object conditions during fMRI and Verbal and Object conditions outside the scanner.

Task performance yielded three outcome variables of interest: reaction time, number of omission errors (failures to respond to a target stimulus), and number of commission errors (responding to a distractor stimulus). Reaction times were averaged across load condition for each participant. Performance accuracy was calculated separately for each modality and load condition using the following formula: $\text{Accuracy} = \frac{\text{Hits} + \text{Correct Rejections}}{\text{Total Stimuli}}$, where $\text{Hits} = \text{Number of Targets} - \text{Omission Errors}$ and $\text{Correct Rejections} = \text{Number of Distractors} - \text{Commission Errors}$. A total of six outcome measures per participant were obtained for each N-back task completed in a specific environment. For example, the Verbal N-back task paradigm completed during fMRI yielded both reaction time and accuracy data for 0-back, 1-back, and 2-back conditions.

Statistical Analyses

All variable distributions were examined for normality in order to determine whether the use of parametric statistics was appropriate. Descriptive statistical analyses were calculated to characterize the sample with regard to demographics and performance on measures of intelligence and working memory. Given the negatively skewed distributions of performance data, reliability of N-back accuracy during fMRI and outside the scanner was compared using nonparametric regression (Generalized Estimated Equation Modeling).

Reaction time was examined separately by task load and modality for all planned analyses. Reaction times were collapsed across load conditions for post-hoc analyses only. This data reduction technique was used based on expectations of similarity regarding the impact of task order on N-back performance, as well as to enhance the stability of means and standard deviations. Reliability of reaction time was examined using a repeated measures analysis of variance (Repeated Measures ANOVA; [Environment (Inside-Outside) $\times$ Modality (Verbal-Object) $\times$ Load (0–1–2-back)]). Appropriate post-hoc comparisons were used to examine significant findings. Validity was assessed using Pearson correlations to examine the relationship between N-back performance during fMRI with Digit Span, self-report ratings on the working memory scale of the BRIEF (BRIEF-WM), and measured intelligence.

Results

Task Performance

Mean accuracy on the N-back was at or above 95% for Verbal and Object N-back tasks performed during fMRI and outside the scanner. Despite near ceiling performance, an examination of effect sizes resulting from mean comparisons across task load suggests that difficulty was effectively manipulated across nearly all task conditions (Table 3). For the Verbal and Object N-back performed outside the scanner, there was no clinically meaningful change in mean accuracy when load was increased from 0-back to 1-back ($d = 0.00$, $d = 0.00$, respectively).

Reaction time data were normally distributed. Effect size analyses revealed that reaction time was increased across nearly all task conditions (Table 4). An exception was seen for performance on the Object N-back during fMRI, as the effect of increased load on reaction

time was minimal ($d = 0.04$). Overall, performance on both the Verbal and Object N-back tasks during fMRI reflected the parametric manipulation of task difficulty.

Reliability

Reliability of performance accuracy was compared during fMRI scanning and outside the scanner using nonparametric multiple regression (Generalized Estimated Equation). N-back accuracy did not significantly differ across environment ($T_W = 0.02$, $p = .88$, $v = .02$); however, accuracy was differentially impacted by increased load in the Verbal and Object conditions ($\chi^2 = 6.92$, $p = .03$, $v = .28$). Post-hoc analyses were conducted to examine this load by modality interaction using Wilcoxon sign-rank tests with data collapsed across environment (Figure 2). Results revealed that mean accuracy was not significantly different between Verbal and Object tasks for 0-back ($Z = -1.78$, $p = .08$, $r = -.27$) or 1-back ($Z = -1.47$, $p = .14$, $r = -.22$) loads; however, accuracy was significantly higher during the Verbal 2-back when compared to the Object 2-back ($Z = -2.44$, $p = .01$, $r = -.37$).

N-back reaction time was compared during fMRI scanning and outside the scanner using Repeated Measures ANOVA (Environment $\times$ Modality $\times$ Load). A significant modality by environment interaction ($F(1, 42) = 6.52$, $p = .01$, $\eta_p^2 = .13$) was explored with post-hoc paired t-tests, using data collapsed across load (Figure 3). Reaction time increased significantly during fMRI, but the increase was not significantly different for the Verbal and Object tasks ($t(43) = 0.82$, $p = .42$, $d = 0.05$). Although reaction times were faster outside of the scanner, the effect was variable, such that participants were slower during the Verbal N-back ($t(43) = 2.86$, $p = .01$, $d = 0.21$).

Additional post-hoc analyses were conducted in order to explore whether observed differences in reaction time across environment could be accounted for by the order of task completion by environment (e.g., initial exposure during fMRI or outside the scanner) or modality (e.g., initial exposure to Verbal or Object N-back). First, group means were compared using one-way ANOVAs, with order by environment entered as a between group factor. There was a significant effect of task order for the Verbal 1-back, Verbal 2-back, Object 1-back, and Object 2-back task conditions ($\eta_p^2 = .10$, $.13$, $.11$, and $.15$, respectively), suggesting that participants who completed the N-back tasks outside of the scanner first were consistently faster during fMRI, relative to those who first completed the task inside the scanner. Next, a second group of one-way ANOVAs was completed, with order by modality entered as a between group factor. Reaction times did not significantly differ by modality.

N-back Task Validity

Correlational analyses were conducted to examine the relationship between clinical measures of working memory (Digit Span, BRIEF-WM) and N-back performance during fMRI scanning. There was no significant relationships between N-back accuracy and LDSF ($|r| = .01 - .29$). A significant but modest correlation was seen between LDSB and Verbal N-back, such that participants with longer backward digit spans (LDSB) were more accurate during the 2-back condition ($r = .36$, $p = .02$). Reaction time for the Verbal N-back was significantly but modestly associated with LDSB, such that participants with longer

backward digit spans were faster during the 2-back condition ($r = -.31, p = .04$). There was no significant associations between self-reported working memory (BRIEF-WM) and N-back accuracy during fMRI ($|r| = .03 - .19$) or N-back reaction time ($|r| = .01 - .13$).

Relationships among N-back performance during fMRI and clinical measures of intelligence (WASI ABIQ, Vocabulary T, and Matrix Reasoning T) were examined with Pearson correlational analyses. Significant relationships were seen between N-back reaction time and estimated intelligence (ABIQ) and between N-back reaction time and vocabulary knowledge (Vocabulary). Specifically, participants with greater estimated intelligence were slower to respond during the Verbal 0-back ($r = .34, p = .03$) and 2-back ($r = .41, p = .006$) conditions, as well as during the Object 1-back ($r = .31, p = .04$) and 2-back ($r = .32, p = .04$) conditions. Participants with greater vocabulary knowledge were slower to respond during the Verbal 0-back ($r = .32, p = .04$) and 2-back ($r = .36, p = .02$) conditions, as well as during the Object 2-back ($r = .30, p = .05$) conditions. There were no significant relationships between N-back accuracy and measured intelligence (ABIQ, Vocabulary T, Matrix Reasoning T; $|r| = .00 - .20$).

Discussion

The current study examined the clinical utility of the N-back as a measure of working memory in a cohort of healthy young adults. We were particularly interested in whether the environment (inside/outside scanner) impacted the reliability of N-back performance. Our results provide mixed support for psychometric properties of the N-back and underscore concerns regarding the clinical utility of this task as a measure of working memory. Further, our findings suggest that in-scanner performance of the N-back task is impacted by multiple factors, thus providing direction to the design and implementation of fMRI studies of neurocognitive functioning.

Hypotheses regarding the reliability of performance across environment were tested using repeated measures designs that accounted for multiple factors that could potentially contribute to variability in performance. There was no evidence for significant differences in N-back accuracy across environment, although accuracy was differentially impacted by increased load in the Verbal and Object conditions, such performance accuracy was more consistent during the Verbal 2-back when compared to the Object 2-back.

Our hypothesis regarding the reliability of N-back across environment was only partially supported. Participants consistently performed faster during N-back tasks outside of the scanner; however, the effect was varied by task modality (participants were slower to complete the Verbal N-back task during scanning) and order of task completion. Specifically, those participants who completed the N-back tasks outside of the scanner first were consistently faster during fMRI, relative to those who first completed the task inside the scanner. Our findings are not completely explained by the difference in response mechanism across environment. Specifically, participants responded via button press outside the scanner and with a pneumatic squeeze ball during fMRI, which takes longer to register a response. It is worth noting that performance was more affected by the scanner environment for those participants who were not provided with the opportunity to complete a full task

paradigm outside of the scanner prior to fMRI. Even though all participants in our study viewed a short presentation that included an introduction to the fMRI scanner, performance still appears to have been disproportionately affected by the lack of opportunity for out of scanner practice. Taken together, these data suggest that factors such as environment, response modality, and order of task presentation (opportunity for practice prior to scanning) affect the stability of reaction time data. These data have implications for fMRI study design, and suggest that acclimation paradigms consisting of actual exposure to all aspects of the MRI scanner (e.g., lying down during task performance in an enclosed space, scanner noise, etc.) may more effectively minimize error variance due to a reaction to the scanner environment (e.g. anxiety).

We found some evidence for the validity of the N-back as clinically meaningful measure of working memory. Participants with better performance on LDSB performed more accurately on the Verbal 2-back during fMRI. In contrast, there were no additional significant associations between N-back performance and clinical measures of working memory. Results further suggest that N-back reaction time, as opposed to accuracy, may be more strongly associated with performance on measures of intelligence. The finding that those participants with *better* performance on measures of intellectual functioning were *slower* to respond during the N-back is puzzling, as it is in the opposite direction of previous reports (Gevins & Smith, 2000; Hockey & Geffen, 2004). It is possible that there was bias toward conservative response in our sample.

Clinicians are encouraged to consider these findings when attempting to interpret behavior as a marker of brain function. More specifically, while our results support the conceptualization of Digit Span performance as an indicator of function within the prefrontal lobe, they do not suggest that similar conclusions be made with regard to self-report data (e.g., BRIEF-WM). Results add to a growing body of literature emphasizing thoughtful consideration of psychometric properties of clinical and research-based measures of working memory (Redick & Lindsey, 2013). Investigators should consider the utility and psychometric properties of measures of neuropsychological function when designing studies.

Limitations and Future Directions

This study is not without limitations. The purpose of this study was to identify a group of healthy young adults with no history of chronic medical or neurological conditions for use as a comparison to a sample of young adult survivors of childhood brain tumors. As such, the young adults recruited for this study generally performed above age expectations on intelligence measures. This may account for the restricted range of performance evidenced in our N-back accuracy data. Future studies should consider the level of task difficulty in relationship to the characteristics of specific groups prior to study design, particularly with studies that aim to assess the efficacy of clinical intervention. Researchers may wish to consider matching task difficulty to expected level of group performance, in order to more accurately compare cohorts on the intended construct.

Despite these ceiling effects, variability in the N-back reaction time data suggest that parametric variation of difficulty was effective. Further, performance on clinical measures of

working memory (e.g. Digit Span and BRIEF) was generally age typical, suggesting that our cohort demonstrated age typical functioning in the cognitive domains of interest. Finally, our findings of significant positive associations between N-back accuracy and performance on clinical measures of working memory are consistent with our hypotheses regarding task validity and support our ability to interpret data from this cohort.

It is possible that the unexplained variance in reaction time data collected inside the scanner could be due to anxiety in response to exposure to the scanning environment. Future studies may consider including more formal measures of state-based anxiety in the design.

Conclusion

Our data provide limited evidence for the clinical utility of the N-back task as a measure of working memory. In particular, we found N-back reaction time during fMRI scanning to be most strongly correlated with measured intelligence, as opposed to clinical measures of working memory. Despite evidence for parametric variation in task difficulty, our ability to explore the relationship between performance accuracy and working memory was somewhat limited by the restricted range of performance in our sample. Nonetheless, our findings of limited relationships between N-back performance accuracy and clinical measures of working memory are consistent with other studies exploring validity of the N-back task outside of the scanner (for review see Redick & Lindsey, 2013).

Our findings join a growing body of literature that has provided mixed support for the clinical utility of the N-back task as a measure of working memory. Many studies have been conducted to examine the patterns of neural activation that are associated with working memory performance; this body of research suggests that the N-back task appears to be a reliable measure for the purpose of eliciting brain activation, though we did not directly assess this in this paper. Rather, our findings suggest that, at the group level, N-back performance inside the scanner appears to be most similar to performance outside of the scanner when participants are offered the opportunity to practice tasks prior to scanning and allowed to acclimate to the scanner environment. Our findings of limited validity between N-back performance and clinical measures of working memory (e.g., Digit Span) suggest that this task may not be measuring the same construct as tools that are routinely employed in clinical neuropsychological assessment. These findings underscore the importance of considering the psychometric properties of experimentally based tasks when interpreting research findings in a clinical context.

Acknowledgments

This work was supported by the National Cancer Institute (St. Jude Cancer Center Support [CORE] under Grant [P30 CA21765] and the American Lebanese Syrian Associated Charities (ALSAC). Research was conducted at St. Jude Children's Research Hospital in Memphis, Tennessee.

References

- Ayr LK, Yeates K, Enrile BG. Arithmetic skills and their cognitive correlates in children with acquired and congenital brain disorder. *Journal of the International Neuropsychological Society: JINS*. 2005; 11:249–262. [PubMed: 15892901]
- Baddeley, A. Working memory. Oxford, England: Oxford University Press; 1986.

- Barch DM, Mathalon DH. Using brain imaging measures in studies of procognitive pharmacologic agents in schizophrenia: Psychometric and quality assurance considerations. *Biological Psychiatry*. 2011; 70:13–18. [PubMed: 21334602]
- Bennett CM, Miller MB. How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*. 2010; 1191:133–155. [PubMed: 20392279]
- Brébion G, Bressan RA, Pilowsky LS, David AS. Depression, avolition, and attention disorders in patients with schizophrenia: Associations with verbal memory efficiency. *The Journal of Neuropsychiatry and Clinical Neurosciences*. 2009; 21:206–215. [PubMed: 19622692]
- Dawson EL, Shear PK, Strakowski SM. Behavior regulation and mood predict social functioning among healthy young adults. *Journal of Clinical and Experimental Neuropsychology*. 2012; 34:297–305. [PubMed: 22225510]
- de Jonge P, de Jong PF. Working memory, intelligence, and reading ability in children. *Personality and Individual Differences*. 1996; 21:1007–1020.
- Friedman NP, Miyake A, Corley RP, Young SE, Defries JC, D J, Hewitt JK. Not all executive functions are related to intelligence. *Psychological Science*. 2006; 17:172–179. [PubMed: 16466426]
- Friedman NP, Miyake A, Young SE, Defries JC, Corley RP, Hewitt JK. Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General*. 2008; 137:201–225. [PubMed: 18473654]
- Fry AF, Hale S. Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science*. 1996; 7:237–241.
- Gaillard WD, Grandin CB, Xu B. Developmental aspects of pediatric fMRI: Considerations for image acquisition, analysis, and interpretation. *NeuroImage*. 2001; 13:239–249. [PubMed: 11162265]
- Garlinghouse MA, Roth RM, Isquith PK, Flashman LA, Saykin AJ. Subjective rating of working memory is associated with frontal lobe volume in schizophrenia. *Schizophrenia Research*. 2010; 120:71–75. [PubMed: 20303715]
- Gevins A, Smith ME. Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. *Cerebral Cortex*. 2000; 10:829–839. [PubMed: 10982744]
- Glover GH, Mueller BA, Turner JA, van Erp TG, Liu TT, Greve DN, Potkin SG. Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. *Journal of Magnetic Resonance Imaging*. 2012; 36:39–54. [PubMed: 22314879]
- Goldman-Rakic PS. Architecture of the prefrontal cortex and the central executive. *Annals of the New York Academy of Sciences*. 1995; 769:71–83. [PubMed: 8595045]
- Grön G. Auditory and visual working memory performance in patients with frontal lobe damage and in schizophrenic patients with low scores on the Wisconsin Card Sorting Test. *Psychiatry Research*. 1998; 80:83–96. [PubMed: 9727966]
- Handwerker DA, Ollinger JM, D'Esposito M. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*. 2004; 21:1639–1651. [PubMed: 15050587]
- Hanten G, Levin HS, Song JX. Working memory and metacognition in sentence comprehension by severely head-injured children: A preliminary study. *Developmental Neuropsychology*. 1999; 16:393–414.
- Hockey A, Geffen G. The concurrent validity and test-retest reliability of a visuospatial working memory task. *Intelligence*. 2004; 32:591–605.
- Jaeggi SM, Buschkuhl M, Perrig WJ, Meier B. The concurrent validity of the N-back task as a working memory measure. *Memory*. 2010; 18:394–412. [PubMed: 20408039]
- Kane MJ, Conway ARA, Miura TK, Colflesh GJH. Working memory, attention control, and the N-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 2007; 33:615–622.
- Lezak, MD. *Neuropsychological Assessment*. 3rd ed.. New York: Oxford University Press; 1995.
- Lueken U, Muehlhan MER, Wittchen HU, Kirschbaum C. Within and between session changes in subjective and neuroendocrine stress parameters during magnetic resonance imaging: A controlled scanner training study. *Psychoneuroendocrinology*. 2012; 37:1299–1308. [PubMed: 22309826]

- Mahone EM, Martin R, Kates WR, Hay T, Horská A. Neuroimaging correlates of parent ratings of working memory in typically developing children. *Journal of the International Neuropsychological Society: JINS*. 2009; 15:31–41. [PubMed: 19128526]
- Miller KM, Price CC, Okun MS, Montijo H, Bowers D. Is the N-back task a valid neuropsychological measure for assessing working memory? *Archives of Clinical Neuropsychology*. 2009; 24:711–717. [PubMed: 19767297]
- Miyake A, Emerson MJ, Friedman NP. Assessment of executive functions in clinical settings: Problems and recommendations. *Seminars in Speech and Language*. 2000; 21:169–183. [PubMed: 10879548]
- Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD. The unity and diversity of executive functions and their contributions to complex “Frontal Lobe” tasks: A latent variable analysis. *Cognitive Psychology*. 2000; 41:49–100. [PubMed: 10945922]
- Nigg JT, Stavro G, Ettenhofer M, Hambrick DZ, Miller T, Henderson JM. Executive functions and ADHD in adults: Evidence for selective effects on ADHD symptom domains. *Journal of Abnormal Psychology*. 2005; 114:706–717. [PubMed: 16351391]
- Oberauer K, Schulze R, Wilhelm O, Süß HM. Working memory and intelligence -- their correlation and their relation: Comment on Ackerman, Bieier, and Boyle (2005). *Psychological Bulletin*. 2005; 131:61–65. [PubMed: 15631551]
- Owen AM, McMillan KM, Laird AR, Bullmore E. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*. 2005; 25:46–59. [PubMed: 15846822]
- Perlstein WM, Carter CS, Noll DC, Cohen JD. Relation of prefrontal cortex dysfunction to working memory and symptoms in schizophrenia. *American Journal of Psychiatry*. 2001; 158:1105–1113. [PubMed: 11431233]
- Petrides M. Functional organization of the human frontal cortex for mnemonic processing. Evidence from neuroimaging studies. *Annals of the New York Academy of Sciences*. 1995; 15:85–96. [PubMed: 8595046]
- Plichta MM, Schwarz AJ, Grimm O, Morgen K, Mier D, Haddad L, Meyer-Lindenberg A. Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. *NeuroImage*. 2012; 60:1746–1758. [PubMed: 22330316]
- Redick TS, Lindsey DRB. Complex span and n-back measures of working memory: a meta-analysis. *Psychonomic Bulletin and Review*. 2013; 20:1102–1113. [PubMed: 23733330]
- Roberts R, Gibson E. Individual differences in sentence memory. *Journal of Psycholinguistic Research*. 2002; 31:573–598. [PubMed: 12599915]
- Roth, RM.; Isquith, PK.; Gioia, G. Behavior Rating Inventory of Executive Function – Adult Version. Lutz, Florida: Psychological Assessment Resources; 2005.
- Roth RM, Lance CE, Isquith PK, Fischer AS, Giancola PR. Confirmatory factor analysis of the behavior rating inventory of executive function-adult version in healthy adults and application to attention-deficit/hyperactivity disorder. *Archives of Clinical Neuropsychology*. 2013; 28:425–434. [PubMed: 23676185]
- Rottschy C, Langner R, Dogan I, Reetz K, Laird AR, Schulz JB, Eickhoff SB. Modelling neural correlates of working memory: A coordinate-based meta-analysis. *NeuroImage*. 2012; 60:830–846. [PubMed: 22178808]
- Salthouse TA, Atkinson TM, Berish DE. Executive functioning as a potential mediator of age-related decline in older adults. *Journal of Experimental Psychology. General*. 2003; 132:566–594. [PubMed: 14640849]
- Shelton JT, Elliott EM, Hill BD, Calamia MR, Gouvier WD. A comparison of laboratory and clinical working memory tests and their prediction of fluid intelligence. *Intelligence*. 2009; 37:283–293. [PubMed: 20161647]
- Smith EE, Jonides J. Neuroimaging analyses of human working memory. *Proceedings of the National Academy of Sciences of the United States of America*. 1998; 95:12061–12068. [PubMed: 9751790]
- Swanson HL. What develops in working memory? A life span perspective. *Developmental Neuropsychology*. 1999; 35:986–1000.

- Szaflarski JP, Holland SK, Jacola LM, Lindsell C, Privitera MD, Szaflarski M. Comprehensive presurgical functional MRI language evaluation in adult patients with epilepsy. *Epilepsy & Behavior*. 2008; 12:74–83. [PubMed: 17964221]
- Van Leeuwen M, Van den Berg SM, Hoekstra RA, Boomsma DI. Endophenotypes for intelligence in children and adolescents. *Intelligence*. 2007; 35:369–380.
- Waiter GD, Deary IJ, Staff RT, Murray AD, Fox HC, Starr JM, Whalley LJ. Exploring possible neural mechanisms of intelligence differences using processing speed and working memory tasks: An fMRI study. *Intelligence*. 2009; 37:199–206.
- Wechsler, D. Wechsler Adult Intelligence Scale – 3rd edition. San Antonio, TX: The Psychological Corporation; 1997.
- Wechsler, D. Wechsler Abbreviated Scale of Intelligence. San Antonio, TX: The Psychological Corporation; 1999.
- Zou P, Li Y, Conklin HM, Mulhern RK, Butler RW, Ogg RJ. Evidence of change in brain activity among childhood cancer survivors participating in a cognitive remediation program. *Archives of Clinical Neuropsychology*. 2012; 27:915–929. [PubMed: 23079152]

** indicates correct response.*

Figure 1.
Verbal and Object N-back

Figure 2.
Accuracy by Task Load

Data collapsed across load condition (0, 1, 2-back)

Figure 3.
Reaction Time by Environment

Table 1

Participant Demographics

	M ± SD	N (%)
Age (years)	24.8 ± 3.52	
Barratt Simplified Measure of Social Status *	51.7 ± 8.17	
Gender		
Male		21 (47)
Female		24 (53)
Race		
Caucasian		43 (96)
African-American		2 (4)
Ethnicity		
Non-Hispanic		43(96)
Spanish/Hispanic/Latino		1(2)
Mexican/Chicano		1(2)

N= 45.

* Scores range from 8 to 66 with higher scores indicative of higher socioeconomic status

Table 2

Performance on Clinical Measures

	M	SD	Median	Range	d ^b
Wechsler Abbreviated Scale of Intelligence ^a					
Abbreviated IQ (SS)	119.90	7.45	121.00	101.00 – 134.00	1.68
Vocabulary (T)	60.64	6.54	61.00	43.00 – 72.00	1.26
Matrix Reasoning (T)	61.84	4.36	63.00	51.00 – 68.00	1.53
Digit Span (Wechsler Adult Intelligence Scale, 3rd Edition) ^a					
Total Scaled Score	12.09	2.81	12.00	6.00 – 18.00	0.24
Longest Digit Span Forward (raw)	7.41	1.02	8.00	5.00 – 9.00	.
Longest Digit Span Backward (raw)	5.93	1.37	6.00	2.00 – 8.00	.
Longest Digit Span Forward (Z)	0.53	0.77	0.94	-1.24 – 1.77	.
Longest Digit Span Backward (Z)	0.57	0.90	0.59	-2.05 – 1.92	.
Behavior Rating Inventory of Executive Function, Adult Self-Report					
Global Executive Composite (T)	47.13	8.40	47.00	35.00 – 66.00	-0.08
Behavior Regulation Index (T)	46.22	7.70	46.00	35.00 – 64.00	-0.11
Metacognition Index (T)	48.13	9.10	46.00	37.00 – 78.00	-0.05
Working Memory (T)	48.73	7.95	49.00	39.00 – 69.00	-0.04

Total sample size = 45. Standard Scores have a mean of 100 and a standard deviation of 15. T-scores have a mean of 50 and a standard deviation of 10. Scaled scores have a mean of 10 and a standard deviation of 3. Z scores have a mean of 0 and a standard deviation of 1.

^a n = 44.

^b Cohen's d statistic derived from mean comparisons between sample data and normative expectations using a pooled standard deviation.

Table 3

N-back Accuracy

	Inside					Outside				
	M	SD	Median	Range	d ^a	M	SD	Median	Range	d ^a
Verbal										
0-back	0.99	0.03	1.00	0.89 – 1.00	.	0.99	0.01	1.00	0.96 – 1.00	.
1-back	0.98	0.03	1.00	0.86 – 1.00	0.33	0.99	0.02	0.99	0.91 – 1.00	0.00
2-back	0.96	0.04	0.98	0.86 – 1.00	0.57	0.96	0.03	0.96	0.90 – 1.00	1.20
Object										
0-back	0.99	0.02	1.00	0.90 – 1.00		0.99	0.02	0.99	0.93 – 1.00	
1-back	0.98	0.03	1.00	0.86 – 1.00	0.40	0.99	0.02	1.00	0.90 – 1.00	0.00
2-back	0.95	0.04	0.96	0.85 – 1.00	0.86	0.95	0.04	0.96	0.85 – 1.00	1.33

N = 44. Inside – performance during fMRI, outside – performance outside of scanner.

^aCohen's d statistic derived from comparisons between 0-back and 1-back, 1-back and 2-back.

Table 4

N-back Reaction Time (ms)

	Inside					Outside				
	M	SD	Median	Range	d ^a	M	SD	Median	Range	d ^a
Verbal										
0-back	5105.00	833.00	4963.00	3796.00 – 7754.00	.	4235.00	572.40	4103.00	3324.00 – 5941.00	.
1-back	5312.00	1113.00	5071.00	3672.00 – 9499.00	-0.21	4410.00	821.70	4239.00	3240.00 – 7404.00	0.25
2-back	6042.00	1425.00	6022.00	3787.00 – 10813.00	-0.58	5410.00	1279.00	5168.00	3043.00 – 8822.00	-0.95
Object										
0-back	5284.00	879.90	5241.00	4054.00 – 8922.00	.	4192.00	460.50	4152.00	3541.00 – 5916.00	.
1-back	5251.00	950.50	5207.00	3720.00 – 8445.00	0.04	4316.00	754.20	4170.00	3434.00 – 6775.00	-0.20
2-back	6052.00	1408.00	5976.00	3363.00 – 9584.00	-0.68	5089.00	1307.00	4733.00	3350.00 – 9693.00	-0.75

N = 44. Inside – performance during fMRI, Outside – performance outside of scanner.

^aCohen's d statistic derived from comparisons between 0-back and 1-back, 1-back and 2-back.