Title

The night out task and scoring application: an ill-structured, open-ended clinic-based test representing cognitive capacities used in everyday situations.

Journal

Archives of Clinical Neuropsychology, 36(4)

Authors

Schmitter-Edgecombe, Maureen

Cunningham, Reanne

McAlister, Courtney

et al.

Publication Date

2021-05-21

DOI

10.1093/arclin/acaa080

Peer reviewed

OXFORD
UNIVERSITY PRESS

Archives of Clinical Neuropsychology 36 (2021) 537–553

Archives of CLINICAL NEUROPSYCHOLOGY

The night out task and scoring application: an ill-structured, open-ended clinic-based test representing cognitive capacities used in everyday situations

Maureen Schmitter-Edgecombe[1], Reanne Cunningham[1], Courtney McAlister[2], Kayela Arrotta[3], Alyssa Weakley[4]

[1]*Department of Psychology, Washington State University, Pullman, WA, USA*
[2]*Mayo Clinic School of Medicine, Mayo Clinic Health System, La Crosse, WI, USA*
[3]*Neurological Institute, Cleveland Clinic, Cleveland, OH, USA*
[4]*Neurology Department, School of Medicine, University of California, Davis, Sacramento, CA, USA*

*Corresponding author at: Department of Psychology, Washington State University, Pullman, Washington 99164-4820, USA. Tel: +509-335-0170. *E-mail address*: schmitter-e@wsu.edu (M. Schmitter-Edgecombe)

Received 17 June 2020; revised 7 August 2020; Accepted 7 September 2020

Abstract

Objective The night out task (NOT) was developed as a naturalistic, open-ended, multitasking measure that requires individuals to complete eight subtasks comparable to those encountered during real-world functioning (e.g., pack travel bag, prepare tea). We examined psychometric properties and administration feasibility of this direct observation measure within a clinic-like setting using a tablet-based coding application.
Method A sample of 148 community-dwelling older adults (82% cognitively healthy; 18% mild cognitive impairment) and 57 younger adults completed the NOT along with other neurocognitive tests and questionnaires.
Results Inter-rater reliability across NOT primary (i.e., time, accuracy, efficiency) and process-related (e.g., error-types, self-corrections) variables was mostly excellent. NOT primary measures showed expected patterns of convergent and discriminant validity with measures of cognition, demographics, and well-being. External validity was established by the NOT ability to distinguish between age and diagnostic (cognitively healthy vs. mild cognitive impairment) groups. Demonstrating incremental validity, the NOT primary variables (execution time in particular) were predictive of self-reported functional abilities and completion quality of in-home everyday tasks over and earlier variables such as demographics, cognition, and mobility.
Conclusions These findings suggest that the NOT and its app interface, which allows for continuous logging of observations, are a feasible in-clinic measure to assess cognitive capacities important for real-world functioning. With further validation, the NOT may allow for earlier detection of functional difficulties. Understanding errors and strategies used during NOT performance could also have implications for individualized interventions.

Keywords: Ecological validity; Multitasking; Everyday functioning; Executive functions; Aging; Mild cognitive impairment

Introduction

Accurately predicting how well older individuals' complete activities of daily living, such as food preparation and medication management, have important implications for clinical recommendations and safety. Although neuropsychologists are often asked to answer questions about real-world functioning, data from several comprehensive reviews indicate that cognition explains an average of 21%–23% of the variance in functional outcomes (McAlister, Schmitter-Edgecombe, & Lamb, 2016; Royall et al., 2007). Executive functions have consistently been found to account for a larger percentage of variance in everyday functioning

doi:10.1093/arclin/acaa080 Advance Access publication 22 October 2020

than other cognitive domains (e.g., attention, speeded processing; McAlister et al., 2016; Royall et al., 2007). Nonetheless, popular executive functioning assessments (e.g., Trail Making Test) only weakly represent cognitive capacities used in everyday situations. For example, few executive functioning tests are ill-structured and open-ended and require multitasking and complex planning abilities (see Chan Shum, Toulopoulou, & Chen, 2008 for a review of executive tests).

To improve the ecological validity of neuropsychological tests, or the extent to which these tasks can represent real-life outcomes (Sbordone, 1996), several approaches to assessment have been advanced. These approaches include creating more naturalistic performance-based tasks (e.g., Baum, Morrison, Hahn, & Edwards, 2003; Schwartz, Buxbuam, Ferraro, Veramonti, & Segal, 2003; Wilson, Alderman, Burgess, Emslie & Evans, 1996), using computerized and virtual reality environments (e.g., Cipresso et al., 2014; Craik & Bialystok, 2006; Parsons & Barnett, 2017; Rand, Rukan, Weiss, & Katz, 2009), and testing individuals in simulated and real-world environments (e.g., Lamberts, Evans, & Spikman, 2010; Rocke, Hays, Edwards, & Berg, 2008; Sanders & Schmitter-Edgecombe, 2012). Each of these approaches has inherent strengths and limitations. For example, task administration in the real-world environment can be time-consuming and costly (Robertson & Schmitter-Edgecombe, 2017). Furthermore, several studies have found that older adults perform more poorly on computerized versions of everyday tasks compared with similar prop-based versions (Kosowicz & MacPherson, 2017; Feinkohl, Cress, & Kimmerle, 2016). In this study, to improve upon the ecological validity of traditional clinic-based tests, we introduce the night out task (NOT) with a tablet-based data collection interface.

The NOT was modeled after naturalistic tasks completed in real-world environments, which have demonstrated higher degrees of ecological validity than most traditional neuropsychological tests (Burgess et al. 2006). For example, the Multiple Errands Test (MET; Shallice & Burgess, 1991) and many of its variants (e.g., MET-simplified version; Alderman, Burgess, Knight & Henman, 2003; MET-hospital version; Knight et al., 2002; MET-Home; Burns et al., 2019) typically give participants an instruction sheet and ask them to complete a number of tasks (e.g., buy a birthday card) in a real-world environment (e.g., shopping mall, hospital, home) while following arbitrary rules (e.g., limit on money that can be spent). These studies have demonstrated associations with functional outcomes, including self-report measures and instrumental activities of daily living (IADL; e.g., Dawson et al., 2005; 2009), whereas traditional executive functioning tests have failed to show such associations (e.g., Bulzacka et al., 2016). Similarly, a study completed with the day out task (DOT; Schmitter-Edgecombe et al., 2012), which was carried out in a campus apartment, demonstrated that the DOT, but not other cognitive tests, was predictive of knowledgeable informant report of everyday functioning in individuals with mild cognitive impairment (MCI; Schmitter-Edgecombe et al., 2012).

Prior studies also reinforce the importance of capturing data other than accuracy and time when predicting real-world functioning, including process-related variables such as the types of errors being made, strategies used, and task efficiency (Alderman et al., 2003; Sanders, Low, & Schmitter-Edgecombe, 2014). For example, on the BMET-R, participants with acquired brain injuries were found to break more rules, commit more task-unrelated inefficiencies, omit more tasks and engage in less strategy use than controls (Clark et al., 2017; Meijer & Krampe, 2018). On everyday tasks that were performed in a campus apartment, younger adults (YA) were more efficient than healthy older adults (HOA), as evidenced by better subtask sequencing and fewer task inefficiency errors (McAlister & Schmitter-Edgecombe, 2013; Schmitter-Edgecombe & Parsey, 2014). Further, HOA made fewer task step omissions and left fewer subtasks incomplete when compared with individuals with MCI (Schmitter-Edgecombe & Parsey, 2014; Schmitter-Edgecombe et al., 2012). These data suggest that the opportunity to capture process-related aspects of performances, like engagement in greater planning or self-monitoring activities, may improve ability to differentiate between diagnostic groups and understanding of the nature of real-life functional limitations.

Building on this prior work, the NOT was designed to assess real-world functioning within the clinic environment. Unlike most traditional executive functioning tasks that examine executive components in isolation, the NOT requires an individual to prepare for a night out by accurately completing and efficiently interweaving eight subtasks (e.g., bring a snack, gather correct change). The NOT story instructions contain several rules (e.g., type of snack, when to make phone call), and the story information is not arranged in an efficient order. The NOT therefore requires individuals to engage multiple cognitive abilities, including multitasking, planning, goal-monitoring, problem-solving, and prospective memory, which are essential in many ill-structured everyday situations (Burgess et al., 2000). The clinic-based administration reduces practical barriers, such as cost, administration time, and the need to norm the task in different environments, which have proven to be challenges for versions of the MET (Castiel et al., 2012). The clinic-based administration also allows for better experimental control over the external situation, whereas the task itself allows participants to apply real-world personal experience to the task problems.

Another unique aspect of the NOT is that the examiner uses a tablet interface to track performance in real time. Although the NOT is completed in the clinic, participants must physically move between locations in the room that contain props (e.g., desk organizer, thermos). The application (app) interface allows the examiner to continuously register data about a participant's actions and errors, as they move around the room interweaving and completing the various subtasks. The tablet-based interface has the advantage of facilitating the evaluation of a participant's approach to completing the NOT and reducing scoring burden. The

interface allows for easy capture of process-related variables that may affect performance (e.g., multitasking, self-monitoring, varying errors types) and provide a more nuanced understanding of each person's difficulties, which may have important implications for individualized interventions.

In this study, we first examine whether the tablet-based interface can be used to accurately capture NOT primary scores (i.e., execution time, accuracy, and efficiency) and process-related variables by examining inter-rater reliability. Secondly, we examine convergent and discriminant validity of the NOT primary variables. Thirdly, we examine incremental validity by assessing whether, in comparison to other cognitive and performance-based tasks, the NOT adds incremental value to predicting everyday functioning as assessed by self-report and by direct observation of individuals in their own home. Fourthly, we examine external validity by evaluating whether the NOT is sensitive to distinguishing performances of YA, HOA, and individuals with MCI. We also investigate process-related variables that may uniquely distinguish between groups. We hypothesize that the NOT will be sensitive to the aging process and distinguish between diagnostic groups (i.e., HOA vs. MCI) given the relatively complex, dynamic nature of the NOT. We further hypothesize that the NOT will associate with complex neurocognitive tasks (i.e., executive functioning and memory) and, compared with traditional neuropsychological tasks, better capture the complexities of real-world performance.

Materials and Methods

Participants

Participants for the primary analyses were 148 community-dwelling older adults who completed one of five different studies in our laboratory between 2015 and 2020. Recruitment methods and screening processes were similar across the studies. Participants were recruited through community health and wellness fairs, advertisements, physician referrals, referrals from local agencies working primarily with older adults, and from past studies in our laboratory. Participants were first screened via phone with a medical screening interview and the Telephone interview for cognitive status (TICS; Brandt & Folstein, 2003). This procedure was used to exclude participants who would have difficulty completing the respective testing battery due to sensory or motor impairment or significant cognitive impairment. All studies were approved by the university institutional review board, and participants received an abbreviated report and an honorarium in return for their participation.

Identical criteria were applied across the five studies to identify those individuals who met criteria for MCI ($n = 26$). However, because the neuropsychological testing protocol varied across studies, the specific tests used to provide evidence of objective cognitive impairment differed across studies. For all studies, an experienced neuropsychologist and graduate students in training determined diagnostic group (i.e., HOA, MCI or dementia) by carefully reviewing clinical interview, neuropsychological testing, and collateral medical information (e.g., results of laboratory and brain imaging data when available). Criteria used to define MCI are consistent with the National Institute on Aging-Alzheimer's Association workgroup (Albert et al., 2011) and include (a) cognitive concern reflecting a change in cognition for at least 6 months; (b) objective evidence of impairment with two or more scores in a single domain falling at least 1.0 standard deviation below age-matched norms (Jak et al., 2009) and in reference to the individual's educational background; (c) mild functional impairment for complex tasks, but basic activities of daily living generally preserved; and (d) not demented. HOA reported experiencing no significant cognitive changes, scored within normal limits on the TICS, and did not meet criteria for MCI. Participants who met criteria for dementia were excluded from this work given the significant difficulties that they experience completing complex everyday tasks.

For the analysis investigating age-effects only, 57 healthy YA (i.e., University students) also served as participants. YA were recruited through the psychology department subject pool and completed the study to fulfill research aspects of course requirements. For all participants, exclusionary criteria included a learning disability, attention deficit hyperactivity disorder (ADHD), English as a second language, and a significant medical (e.g., stroke, Parkinson's disease) or psychiatric history (e.g., alcohol/drug abuse), as these deficits may differentially affect NOT performance, and we did not have large enough samples to separately evaluate.

Measures

The following cognitive, functional, and psychosocial measures were used in the varying analyses to examine the psychometric properties of the NOT. Given that the NOT data came from several different studies, not all participants had data for each measure.

Telephone interview for cognitive status. The total score from this brief, telephone-based mental status exam was used as a screening measure of global cognitive functioning (Brandt & Folstein, 2003).

Trail Making Test. Participants rapidly alternated between connecting numbers (Trail Making Test [TMT]-A) and numbers and letters (TMT-B; Reitan & Wolfson, 1985). The time to completion for TMT-A and for TMT-B were used as measures of processing speed and executive function, respectively.

Delis–Kaplan executive function system letter fluency and category subtests. Participants had 60 seconds to quickly produce words beginning with specified letters and that belonged to specified categories (Delis et al., 2001). The total correct words produced for the letters F, A, and S, and the categories Animals and Boys' Names were used as measures of executive function and language, respectively.

Activity-based multiple memory processes paradigm. Before completing 10 different laboratory activities, each taking 5–12 min (e.g., TMT), participants were told that their memory for the tasks and the order of task completion would be tested. Participants were also presented with a task challenge rating card after each task and told to use the card and their response as a cue to remember to ask the examiner for a pill bottle. Once the first task started, no additional reference to task instructions was made (Schmitter-Edgecombe, Woo, & Greeley, 2009). The number of times (10 maximum) the participant correctly reminded the examiner to retrieve the pill bottle was used as a measure of prospective memory. Recall was represented by the number of 10 activities correctly described, and temporal order memory by the absolute deviation between the correct order of activity completion and the order produced by the participant.

Rey auditory verbal learning test. Participants were read an unrelated list of 15 words five times and asked to recall the list of words after each presentation. Following 20 min of distractor tasks, participants were again asked to recall the word list (Rey, 1964). The total number of words recalled across trials 1–5 was used as a measure of list learning and the number of words recalled after the 20-minute delay as a measure of delayed memory.

Weschler adult intelligence scale-IV, digit span forward subtest. Participants were read a sequence of numbers that increased in length and repeated the numbers back to the examiner in the same order (Wechsler, 2008). The total number of correct trials was used as a measure of attention.

Timed up and go test. Participants stood up from a seated position, walked 10 feet, turned around, walked back to a chair, and sat down (Podsiadlo & Richardson, 1991). The mobility score was represented by the total time it took the participant to complete the sequence.

Patient-reported measurement information system, sleep disturbance, and emotional distress short forms (eight items). These measures were used to assess psychological well-being over the past 7 days (1–5 Likert responses). The sleep disturbance scale assessed self-reported perceptions of sleep quality, depth and restoration associated with sleep, and the emotional distress scale assessed aspects of anxiety. Higher scores represent more of the concept being measured (Yu et al., 2011; Pikonis et al., 2011).

Instrumental activities of daily living-compensation. Individuals self-reported their ability to complete IADLs on a scale from 1 (independent; as well as ever, no aid) to 8 (not able to complete the activity anymore). The participant could also indicate if the activity was always completed by someone else. The total score is the mean of items (27 total) responded to using the eight-point scale (Schmitter-Edgecombe, Parsey, & Lamb, 2014).

Performance-based score and observed quality score. Participants were administered a battery of validated performance-based subtests (e.g., Independent Living Scale, Loeb, 1996; Observed Test of Daily Living-Revised, Goverover & Josman, 2004) and a real-world assessment of 10 semi-structured everyday activities (e.g., prepare a light lunch, change a light bulb, plan an event for tomorrow) in their own homes. The performance-based tasks were selected for their likeness to the real-world tasks that participants completed. See Weakley, Weakley and Schmitter-Edgecombe (2019) for details of how the performance-based score was derived and how the quality score was computed based on participant performance on the real-world task. Higher performance-based and quality scores indicate better performance.

Fig. 1. Panoramic view of example NOT testing environment.
Note. Participant is seated at the table in the middle chair with the organizer in front of them. Cupboard A is to the left of the participant's seat. Cupboard B, the Hot Water Pitcher and the Timer are to the right. Travel bag hanging on the door behind the participant.

Night out task. The majority of participants (82%) completed the NOT in the same laboratory room. The laboratory testing environment was a 15 × 25-foot room; the NOT can be administered in a smaller testing room and within a participant's home as it was for some participants in this sample. Props were placed around the room in five standard locations, and participants walked between locations throughout the duration of the task (see Fig. 1). Participants were told to imagine that they were planning to leave for a night out, which would include meeting a friend at the theater for the last showing of a movie and later traveling to the friend's house for dessert. The eight subtasks that needed to be completed to prepare for the night out were then clearly explained (see Appendix A for instructions), and participants were provided a detailed written list of the subtask requirements (see Table 1). However, the provided information was not arranged in an efficient order as some subtasks should be interweaved (e.g., tea must steep for 3 min), whereas other subtasks must be completed in a specific order (e.g., locate movie schedule before gathering money) or at a specified time (e.g., make phone call just before leaving home). Participants were given the opportunity to ask questions. Before beginning the NOT, participants were reminded to multitask and interweave tasks so that the tasks could be completed efficiently (Schmitter-Edgecombe & Cunningham, 2019).

As can be seen in Fig. 2, all eight activities are displayed on the tablet screen along with their respective possible errors (e.g., takes more than one snack) and a start/stop and comment button. There are also buttons for the examiner to log double-checking, self-correction, nontask-related activities, and mid-task planning behaviors. As the individual completes the activities, the examiner observes and selects the appropriate start/stop buttons to designate when the participant is actively engaged in each of the tasks. As individuals are encouraged to multitask and interweave tasks, the start/stop buttons for each task are pressed multiple times. The examiner also records errors and other behaviors such as mid-task planning by pressing the appropriate button at the time the behavior is occurring. A summary page at the end allows the examiner to enter additional information needed for calculation of NOT scores (e.g., time participant wrote down they would leave the house). Significant pilot work was conducted in order to generate the list of potential errors that are displayed on the app and designated as either inefficient or inaccurate/incomplete error types.

Later, two coders who were masked to the study hypotheses, watched video data, and used the NOT app to score participant performances. Table 2 and Appendix B provide detailed information about how each of the NOT variables used in this study was computed. First, three primary variables representing different dimensions of complex everyday behaviors were derived: execution time (lower is better), accuracy (lower is better), and efficiency (higher is better). Errors were coded as inefficient or inaccurate/incomplete, and self-monitoring behaviors were represented by double-checking and self-corrections. Preplanning and proportion of time spent mid-task planning were calculated, and metrics were also derived to represent multitasking behaviors and task interruptions.

Analyses

Because demographically adjusted normalized scores were not available for all tests (cognitive, mobility, and questionnaires), raw scores were used in the analyses. Several variables were normalized either by trimming significant outliers (NOT execution time; the data from 1 YA and 3 HOA was replaced with a NOT execution time that was three standard deviations from the

Table 1. List of goals given to participants

Night Out Task: Things to do list
Objective: Be at the movie theater 10 minutes before the movie, *Muppets Most Wanted,* starts to meet friend
Choose a movie snack to share (Cupboard A). Your friend loves milk chocolate but does not like dark chocolate.
Just prior to leaving the house, call your friend and tell them that you are leaving. Their phone number is 555-1234 (Organizer on Table).
Gather ingredients for Puppy Chow Bars (Cupboard A and Cupboard B).
Plan trip to movie theater; record movie start time, time you must leave your house, and total cost of movies (Organizer on Table).
Bring travel bag to front door and walk through the door.
Prepare your tea for 3 minutes (set the timer) using the water pitcher, teabag and thermos (Cupboard A).
Get correct change for movie tickets (Organizer on Table).
Pack everything into travel bag (Hanging).
Locate Puppy Chow Bars recipe in recipe book (Cupboard B).
Cost of movie tickets for you and your same aged friend: _____
Movie start time: _____
Time must leave house to get to the movie theater 10 minutes before the last showing of the movie, *Muppets Most Wanted* (note: It takes 15 minutes to get to the movie theater): _____
Note: The activities can be completed in any order. Please multitask and interweave to Complete the tasks in an efficient and natural way.

mean of the respective group) or by applying an appropriate transformation that resulted in reducing skewness and kurtosis of the distribution to < 1.2 for the respective variable. Reciprocal square root transformations were used for TMT-B, IADL-compensation (IADL-C), and timed up and go (TUG), and a log10 transformation for the NOT preplanning and proportion mid-task planning variables. Intraclass correlation (ICC) coefficients were used to examine inter-rater reliability for NOT primary and process-related variables. Pearson or Spearman correlational analyses were conducted between the NOT primary variables and variables representing demographics, cognition, mobility, and well-being to examine for convergent and discriminant validity. Due to the number of correlations conducted, a more conservative p-value of $p < .01$ was used for significance. To demonstrate ecological validity, a hierarchical linear regression analysis was conducted to determine whether NOT primary variables would

Fig. 2. NOT app screen shot before testing begins.
Note. Light gray error buttons indicated inefficiency errors, and dark gray error buttons indicated inaccurate/incomplete errors.

account for significant variance in self-reported everyday functioning after accounting for demographics, mobility, and cognition. A hierarchical regression was also used to examine the ability of the NOT to account for significant variance in quality of observed everyday activity completion (quality score) after accounting for performance-based tests. To determine whether the NOT showed external validity, we conducted one-way analyses of variance (ANOVA) comparing YA, HOA, and MCI groups using the Welch statistic and Games-Howell post hoc tests because the assumption of homogeneity of variance was violated for most comparisons. Finally, we used Pearson correlation coefficients to examine test–retest reliability for the NOT primary variables.

Results

Unless otherwise stated, participants were 148 community-dwelling adults age 50+ (M age = 68.43; SD = 8.56; range 50–90; 69% female), with an average education level of 16.31 years (SD = 2.34; range 12–20). Twenty-six of these individuals met criteria for MCI (n = 26; M age = 72.85, SD = 6.10; M education = 15.92, SD = 2.50; 62% female).

Inter-rater Reliability

Inter-rater reliability was calculated for 20% of the sample. The ICC and 95% confidence intervals for the NOT primary variables and process-related variables are displayed in Table 2. With the exception of double-checking behaviors (ICC = 0.80), inter-rater reliability was found to be excellent for all NOT scores with ICCs > 0.90.

Correlations Among NOT Primary Variables

Pearson correlations revealed low but significant correlations between NOT execution time and both the NOT accuracy ($r = 0.29$, $p < .001$) and efficiency ($r = -0.24$, $p = .003$) scores. NOT accuracy and efficiency were moderately correlated ($r = -0.47$, $p < .001$), such that higher NOT performance efficiency (i.e., task sequencing) was associated with making fewer errors and more accurate performance.

Table 2. Night out task (NOT) variables with descriptors and inter-rater reliability

NOT variable	ICC	95% CI	Description
NOT primary variables			
Execution time	0.99	0.99–1.00	Amount of time required to complete the NOT; excludes preplanning time.
Accuracy	0.98	0.95–0.99	Each of the eight subtasks is assigned a task completion score (see Appendix B) and these scores are summed; range 8 (best)–32 (worst).
Efficiency	0.97	0.93–0.98	Total number of eight activities that are sequenced correctly (e.g., recipe read before retrieving food items; See Appendix B).
NOT process-related variables			
Inefficient errors	0.96	0.92–0.98	Coded when an action slows down (e.g., looks in multiple locations) or compromises the efficiency of task (waits for tea, not multitasking) but the task can still be completed.
Inaccurate/Omission errors	0.97	0.94–0.99	Coded when a step or subtask necessary for accurate NOT task completion is not performed (e.g., gets phone but does not make call) or is performed inaccurately (e.g., records incorrect time; makes coffee not tea) or is not attempted.
Double-checking	0.80	0.60–0.90	Coded each time a participant goes over something again, such as going back over the ingredients list after collecting the baking items or going back over the NOT list after finishing multiple subtasks.
Self-corrections	0.92	0.84–0.96	Coded when a participant completes a task or part of a task incorrectly but then fixes their own error.
Nontask-related activity	Low base rate		Coded when a participant engages in something unrelated to the NOT.
Preplanning time	0.99	0.99–1.00	Coded as amount of time elapsed from end of instructions until participant begins to engage with the first task.
Proportion time mid-task planning	0.92	0.83–0.96	Represents proportion of execution time that participant spent engaged in planning after the NOT started.
Multitasking	0.99	0.97–0.99	Represents number of instances two or more tasks overlap in total time; the tasks do not need to actively be completed at the same time (range 1–56)
Interruptions	0.93	0.85–0.96	Represents how many times a participant stopped a task and came back to it later.

Note: ICC = intraclass correlation coefficient; CI = confidence interval. See also Fig. 2 and Appendix B.

Convergent and Discriminant Validity

Pearson or Spearman (for gender) coefficients used to examine convergent and discriminant validity are reported in Table 3. We considered low correlations (0.0–0.3) to indicate discriminant validity and moderate correlations (0.3–0.6) to represent convergent validity. Sample sizes are also reported in Table 3 as different measures were administered to the 148 community-dwelling adults across studies. To improve interpretation, the sign of the correlations was transformed such that positive correlations mean that both measures represent better performance or more positive well-being. Consistent with expectations, low correlations (0.0–0.3) were observed between the NOT measures and gender, education, mobility, and psychosocial wellbeing. Although correlations were low, better mobility was associated with faster NOT execution time ($r = 0.29$), and higher levels of education were associated with better NOT accuracy and efficiency scores ($rs = 0.26$). As expected, age correlated significantly with the primary NOT variables, such that higher age was associated with poorer performances. Also consistent with expectations, the cognitive measures that assessed global cognitive status, memory, and executive function generally showed moderate correlations with the NOT primary measures (0.3–0.6), whereas measures that assessed attention (Digit Span Forward) and language (Category Fluency) showed no significant correlations ($rs \leq 0.22$). The one exception was the finding that better memory scores ($rs = 0.28$–0.36) but not executive function measures ($rs = 0.18$–0.24) were related to higher NOT efficiency. Furthermore, a measure of speeded processing (TMT-A) correlated significantly with NOT execution time ($r = 0.43$) but not with NOT efficiency or accuracy.

Incremental Validity: Self-Reported IADL Performance

Of the 148 community-dwelling participants, 117 completed the self-report version of the IADL-C (higher score = better performance due to normality transformation). This subsample was 66% female, with an average age of 68.27 years ($SD = 8.86$;

Table 3. Pearson coefficients between night out task (NOT) scores and measures of cognition, mobility, psychosocial wellbeing, and demographics

	Primary main scores		
	NOT execution[a]	NOT efficiency	NOT accuracy[a]
Global cognitive status			
TICS, $n = 148$	.37**	.25*	.35**
Executive			
Trail Making Test B, $n = 71$	.52**	.18	.37*
D-KEFS letter fluency, $n = 113$	.26*	.20	.25*
AMMP prospective memory, $n = 74$	.37**	.24	.30*
AMMP temporal order memory[a], $n = 74$	.48**	.21	.25
Memory			
AMMP recall, $n = 74$	.42**	.36**	.34*
RAVLT list learning, $n = 74$	.55**	.36**	.33*
RAVLT delayed memory, $n = 74$	.60**	.28*	.19
Speeded processing/attention			
Trail Making Test A[a], $n = 71$	.43**	-.02	.18
WAIS-IV digit span forward, $n = 113$	.14	-.07	.19
Language			
D-KEFS category fluency, $n = 113$	.22	.09	.10
Mobility			
Timed up and go test, $n = 147$	.29**	.16	.15
Psychosocial wellbeing			
PROMIS sleep disturbance[a], $n = 102$	.05	-.06	-.11
PROMIS distress[a], $n = 57$	.06	-.16	-.08
Demographics			
Gender, $n = 148$[b]	.15	.13	-.14
Education, $n = 148$	.01	.26*	.26*
Age, $n = 148$	-.44**	-.31**	-.28**

Note: A positive correlation means that performance on both measures reflects better performance or more positive well-being. TICS = telephone interview of cognitive status; D-KEFS = Delis–Kaplan executive functioning scale; AMMP = activity memory paradigm; RAVLT = Rey auditory verbal learning test; WAIS-IV = Wechsler adult intelligence scale; PROMIS = patient reported measurement information system.

[a]Reversed in table (if not already reversed by transformation) such that higher score means better performance or more positive well-being.

[b]Spearman coefficient calculated (1 = female; 2 = male).

**$p < .001$;

*$p < .01$.

range = 50–90) and an education of 16.26 years ($SD = 2.34$; range 12–20); 20% of the sample met criteria for MCI. In block one of the hierarchical regression analysis, we entered the demographic variables of age and education, which correlated with NOT performances. In block 2, we entered measures of cognition (Telephone interview for cognitive status [TICS]) and mobility (TUG), which correlated with NOT performances and were available for all study participants. In block three, we entered the NOT primary variables (execution time, efficiency, accuracy). Variance inflation factor values were acceptable (<1.52). Data from the regression analysis are displayed in Table 4. After accounting for the demographic variables in block 1, which did not account for significant variance in the IADL-C score, $F(2,114) = 0.24$, $p = .79$, the measures of cognition and mobility explained a significant additional 6% of the variance, $F(2,112) = 3.44$, $p = .04$. When the NOT primary variables were entered in block 3, a significant additional 24% of the variance in the IADL-C score was accounted for, $\Delta F(3,109) = 11.30$, $p < .001$. In model 2, after controlling for all other variables, the cognitive measure ($t = 2.32$, $p < .05$) emerged as the only significant predictor of the self-reported IADL-C score. In the final model, both age ($t = 3.00$, $p = .003$) and NOT execution time ($t = -5.78$, $p < .001$) emerged as significant predictors of self-reported everyday functioning.

Incremental Validity: In-home Performance Assessment

Incremental validity was also evaluated in a small subsample of community-dwelling older adults ($N = 18$) who completed the NOT, a battery of validated performance-based subtests and a real-world assessment in their own homes. Seventy-eight percent of the sample was female, with an average age of 71.56 years ($SD = 10.52$, range = 55–90) and education of 15.59 years ($SD = 2.50$, range 12–20). In this small subsample, correlations among the NOT variables were similar in pattern to the larger sample, with a moderate correlation again found between the NOT accuracy and efficiency scores ($r = -0.41$). Correlational analysis also revealed that faster NOT execution time ($r = -0.50$) and a higher performance-based score ($r = 0.38$) were

Table 4. Summary of hierarchical regression for predictors of self-reported instrumental activities of daily living-compensation scale

Predictors	Model 1	Model 2	Model 3
Age	.06	.15	.30*
Education	.04	-.00	.05
TICS total		.22*	.16
TUG (transformed)		.12	.02
NOT accuracy[a]			.08
NOT efficiency			-.12
NOT execution time[a]			-.55**
Change in R^2	.00	.06*	.22**
Total R^2	.00	.06*	.28**

Note: Standardized beta coefficients presented for predictors. TICS = telephone interview of cognitive status; TUG = time up and go test; NOT = night out task.
[a]Higher scores represented poorer performances.
*$p < .05$.
**$p < .001$.

both moderately associated with a higher quality score on the in-home everyday tasks, although the NOT execution time and performance-based scores were not associated with each other ($r = -0.05$). Due to limited sample size, a hierarchical regression predicting the in-home quality score was conducted entering only the performance-based score in block 1 and NOT execution time in block 2, as NOT accuracy ($r = -0.04$) and efficiency ($r = 0.20$) were not strongly related to the quality score. NOT execution time accounted for a significant amount of variance (23%) in explaining the in-home quality score, $\Delta F(1,15) = 5.43$, $p < .05$, over and above the variance accounted for (14%) by the performance-based score, $F(1,16) = 2.73$, $p = .11$. In the final model, which accounted for 37% of the variance, NOT execution time ($t = -2.34$, $p < 0.05$) was a significant predictor of the in-home quality score (performance-based score: Model 2: $t = 1.74$, $p = 0.10$; Model 1: $t = 1.65$, $p = 0.12$). Of note, the in-home quality score did not involve a time component.

External Validity: Relationship to Age and Clinical Diagnosis

External validity was evaluated by examining differences between a sample of YA, HOA, and individuals with MCI on the NOT primary variables and process-related variables using one-way ANOVA. The YA were college students ($n = 57$; age 18–22). Individuals in the MCI group ($n = 26$) ranged between the ages of 63 and 83, and all HOA that fell within the same age range of 63–83 ($n = 79$) from the full sample were selected for this analysis (see Table 5 for demographics). Nontask-related activities did not occur in the YA sample and had very low base rates in the HOA ($n = 3$; 3.8%) and MCI ($n = 2$; 7.7%) groups, and they were not analyzed.

As seen in Table 5, despite the higher education of the HOA groups, the YA performed better than the HOA, who performed better than the MCI group on the NOT primary variables of execution time and accuracy. For NOT efficiency, the YA and HOA both performed significantly better than the participants with MCI. The examination of the NOT process-related variables revealed that the proportion of time spent in mid-task planning, and the number of inefficient errors and self-corrections were fewest in YA, significantly higher in HOA, and significantly higher yet in the MCI group. Compared with both YA and HOA, individuals with MCI engaged in more preplanning and committed more errors that represented subtasks being completed inaccurately, left incomplete, or not attempted. In comparison to YA, both the HOA and MCI groups engaged in more double-checking behavior. The groups did not differ significantly in multitasking behavior or in the number of task interruptions.

Discussion

There is a need for new clinic-based measures that are more congruent with questions clinicians are asked to address within the context of neuropsychological evaluations (e.g., Can patient live independently?). Unlike many traditional neurocognitive tests, which are structured, paper-pencil tasks with a clear path to performance completion (Chan et al., 2008), the NOT is an example of an ill-structured, multitasking paradigm that requires participants to complete tasks that are more comparable to those encountered in the real-world. We found that it is feasible to administer the NOT within a clinic-like setting to both older and YA populations. The data demonstrated excellent inter-rater reliability across nearly all NOT variables, convergent and discriminant validity of the NOT primary measures, and external validity by capturing aging effects and the impact of mild cognitive impairment (i.e., HOA and MCI). The data also provide some preliminary evidence for the ecological validity of the NOT. That is, after accounting for demographics and other testing data, the NOT added incremental variance (23%–24%) when predicting everyday functioning as measured by self-report and by the quality of observed in-home performances.

Table 5. Mean night out task (NOT) performances (*SD* in parentheses) as a function of age and diagnostic group

Variable	Group			Statistic	
	YA ($n = 57$)	HOA ($n = 79$)	MCI ($n = 26$)	Welch statistic	Games-Howell post hoc
Demographics					
Age	19.19 (1.08)	70.53 (5.27)	72.85 (6.10)	4382.53**	YA < HOA = MCI
Education	13.05 (1.09)	16.51 (2.37)	15.92 (2.50)	72.04**	YA < HOA = MCI
Gender (% female)	63%	69%	62%		
NOT primary variables					
Execution time	474.35 (119.79)	657.58 (165.24)	956.15 (273.30)	54.61**	YA < HOA < MCI
Accuracy[a]	10.49 (1.54)	11.59 (2.51)	13.65 (2.38)	20.53**	YA < HOA < MCI
Efficiency	4.84 (0.94)	4.71 (0.91)	4.08 (1.09)	4.80*	YA = HOA < MCI
NOT process variables					
Inefficient errors	1.89 (1.39)	2.72 (1.97)	4.77 (2.64)	15.06**	YA < HOA < MCI
Inaccurate/Omission errors[b]	0.65 (0.95)	1.01 (1.26)	2.00 (1.52)	8.95**	YA = HOA < MCI
Double-checking	0.51 (0.65)	1.75 (1.49)	2.08 (1.76)	27.77**	YA < HOA = MCI
Self-corrections	0.25 (0.54)	.56 (0.76)	1.04 (0.87)	10.59**	YA < HOA < MCI
Preplanning time[c]	17.96 (19.87)	20.66 (20.07)	38.12 29.41)	9.06**	YA = HOA < MCI
Proportion mid-task planning time[c]	0.06 (0.04)	0.14 (0.09)	0.22 (0.13)	40.37**	YA < HOA < MCI
Multitasking	30.28 (6.90)	28.87 (8.05)	27.11 (8.93)	2.20	ns
Task interruptions	9.77 (3.31)	11.16 (4.23)	11.11 (5.07)	2.50	ns

Note: YA = younger adult; HOA = healthy older adult; MCI = mild cognitive impairment.
[a]Higher accuracy score represents poorer performance.
[b]Represents a total of inaccurate and incomplete errors types and tasks not attempted.
[c]Data presented as not transformed for ease of interpretation.
**$p < .001$;
*$p < .01$.

Consistent with other studies that have demonstrated that naturalistic direct observation measures can have objective and reliable coding systems (e.g., Schwartz et al., 2003; Schmitter-Edgecombe et al., 2012), we found excellent inter-rater reliabilities between coders for almost all NOT scores. One advantage of the NOT app is that it allows for continuous logging of data observations in real time as participants complete the task (e.g., specific errors, each time a task is started and stopped; see Fig. 2). These data are then automatically summed to produce the NOT primary variables, process-related variables, and other raw data statistics. This significantly improves the ease of coding process-related variables when compared with prior naturalistic coding methods (e.g., Dawson et al., 2009). Of note, the data used for establishing inter-rater reliability for this study were scored offline by coders reviewing NOT task videos. Future studies will be needed to determine how much practice is necessary for clinicians to accurately code the NOT in real time. The experiences of our examiners indicate that with practice, accurate scoring can be achieved.

The NOT primary variables were chosen theoretically (Burgess, 1997) to represent functions important for everyday activity completion and included execution time, accuracy, and efficiency (i.e., sequencing). Although accuracy and efficiency are often related, an individual could complete a task accurately (e.g., retrieve all items on a list to begin a project) but not efficiently (e.g., visit the same storage closet multiple times to retrieve items). Correlational analyses revealed that task accuracy and efficiency, which both characterize different types of NOT errors, were moderately correlated with each other. However, NOT execution time showed only low correlations with the measures of accuracy and efficiency. Demonstrating convergent validity, the NOT primary scores generally showed moderate relationships with neurocognitive measures assessing global cognition and higher-order processes of memory and executive functioning. Interestingly, NOT efficiency correlated significantly with the memory measures but not with the executive functioning measures. Prior research with the DOT found stronger associations between the task efficiency score (i.e., sequencing) and prospective memory (executive) as compared with content memory (Schmitter-Edgecombe et al., 2012; McAlister & Schmitter-Edgecombe, 2013). One possibility for the contrasting study findings is that the DOT was performed in a realistic everyday environment (i.e., campus apartment), which may have helped to reduce memory load and supported performance through more naturalistic environmental cues.

Demonstrating divergent validity, there were no significant correlations between the NOT primary measures and gender or self-report measures of well-being. In addition, measures of lower level cognitive processes, including attention and language, exhibited no significant correlations with the NOT primary measures. As might be expected, measures of mobility (TUG) and speeded processing (TMT-A) were found to associate with the NOT execution time measure but not with NOT accuracy or efficiency. Overall, the pattern of relationships among the NOT measures and their associations with other cognitive domains

further supports the theoretical distinction between the NOT primary measures and provides evidence for convergent and discriminant validity.

Demonstrating external validity, the NOT primary and process-related variables were found to be sensitive to the impact of age and the effects of MCI. Both the HOA and MCI groups performed more poorly than the YA on the NOT primary scores assessing execution time and accuracy. In addition, the MCI group demonstrated poorer NOT execution time and accuracy compared with the HOA group. Unlike prior work with the DOT where the efficiency score was found to be poorer for both HOA and MCI compared with YA (Schmitter et al., 2012; McAlister and Schmitter-Edgecombe, 2013), in the current study, NOT efficiency was poorer for the MCI group compared with both YA and HOA. As mentioned earlier, this may reflect differences between the type of environments the two tasks were performed in (i.e., naturalistic vs. lab) or differences between the sequencing demands imposed by the NOT compared with the DOT. The current findings are consistent with prior work with a naturalistic open-ended planning task (i.e., Apartment Map Task or Amap), which found that individuals with MCI exhibited poorer efficiency and accuracy compared with HOA in both the formulation and execution stages of a naturalistic task completed in a campus apartment (Sanders et al., 2014).

An evaluation of error types revealed that the HOA differed from the YA only in inefficient errors. In contrast, the MCI group committed both more inefficient and inaccurate/omission errors compared with both the HOA and YA. This is consistent with prior naturalistic work conducted in a campus apartment, which found that, in the progression from healthy aging to MCI, task difficulties evolved from task inefficiencies to task omission errors, which resulted in inaccurate task completion (Schmitter-Edgecombe & Parsey, 2014). Prior work has also linked task omission errors with impairment in episodic memory (Giovannetti et al., 2008) and with a reduction in hippocampal and medial temporal lobe volume (Bailey et al., 2013; Seidel et al., 2013).

In addition to task accuracy, time, efficiency, and error types differentiating between group performances, the NOT process-related variables may provide insight into different strategic approaches to NOT completion. Compared with the YA and HOA, individuals with MCI spent more time engaging in preplanning and mid-task planning behaviors and exhibited more self-corrections. Despite these adjustments, the MCI group committed more errors of all types and were less efficient during NOT task performance. Compared with the YA, the HOA also engaged in a greater proportion of mid-task planning and made more self-corrections than the YA during NOT task completion. These may represent compensatory behaviors that served to successfully limit the number of sequencing and inaccurate/omission errors for HOA, but not inefficiency errors. Of interest, the amount of multitasking was highest in the YA and lowest in individuals with MCI, suggesting that reducing multitasking behavior may be another strategy that participants are using to assist with managing the complexities of the NOT. Further research will be needed to explore potential implications of these process variables. Recent work suggests that compensatory strategy use can mitigate the effects of cognitive decline on everyday task performances in community-dwelling older adults (Tomaszewski Farias et al., 2020; Weakley et al., 2019).

Understanding the specific types of errors that participants are committing and strategies being used during NOT performance could have implications for the types of rehabilitation techniques that may be most beneficial. For example, participants who experience difficulty efficiently sequencing tasks might benefit from strategies such as Goal Management Training or other strategies that teach planning skills (e.g., Levine et al., 2000; Manly et al., 2002; Turner et al., 2020). Individuals who leave subtasks incomplete or have difficulty with internal self-monitoring might benefit from the use of external compensatory strategies (e.g., checklists, cueing systems) that would assist them with making sure that all aspects of the task had been accurately completed. Tests like the NOT could also make it easier for clinicians to help families better understand the difficulties being experienced by their loved one with cognitive impairment and to put in place strategies that have the greatest potential to assist with improving everyday functioning. Future work with other populations will also be needed to determine whether the NOT pattern of behavioral performances can distinguish between diagnostic groups. For example, one might hypothesize that individuals with ADHD may engage in a greater number of NOT task interruptions than other populations. In addition, similar to our prior work with the DOT (Dawadi et al., 2013; Schmitter-Edgecombe et al., 2012), the NOT is expected to be too complex for participants with dementia to complete. This is consistent with the criteria that individuals with dementia exhibit difficulties completing tasks of daily living and future work will be needed to determine if those who struggle to complete the NOT meet criteria for dementia.

Consistent with the goal of developing a clinic-based test that is more predictive of real-world functioning than traditional measures, we provide preliminary data demonstrating incremental value of the NOT to the prediction of real-world functioning. Given that there is no gold standard for capturing real-world functioning (Gold, 2012), we present data that used both self-report and naturalistic in-home assessment of an individual's ability to complete IADLs as measures of everyday functioning. Both analyses were consistent in demonstrating that the NOT, and in particular the task execution time variable, added significant variance (incremental validity) to the prediction of everyday performance even after accounting for demographics, mobility and, cognition (i.e., global cognitive status screen) in the case of self-reported IADLs, and after accounting for scores on performance-based tests in the case of the in-home assessment. In both cases, an additional 23%–24% of the variance was explained by the

NOT. This preliminary work suggests that not only is the NOT capturing aspects of everyday functioning but also that the NOT is able to add value to the prediction of everyday functioning over and above that of standard cognitive and performance-based clinic measures and demographics.

Future work will be needed to advance normative data for the NOT. Given that the NOT is being developed to predict real-life everyday abilities, one challenge for this work is determining how best to capture real-world functional status. Current proxy measures (e.g., self- and informant-based questionnaires, performance-based measures) each have strengths and weaknesses (e.g., Marson & Hebert, 2006; Dassel & Schmitt, 2008) and do not typically correlate well with each other (e.g., Burton et al., 2009; Finlayson et al., 2003), suggesting that they may capture different aspects of everyday functioning (Schmitter-Edgecombe et al., 2011). Directly observing people in their own homes, as we did for a small subset of study participants, is time-consuming and costly and requires careful coding. Newer assessment methods, such as ambulatory assessment or the use of ambient or wearable sensors that allow for continuous data collection in real-world environments, may offer new opportunities for capturing the quality of real-world performances. Another challenge is determining when demographic variables (e.g., age, education) should be taken into account when generating normative data. At some point if a certain threshold of poor performance is reached on the NOT, this threshold may be predictive of impaired real-world performance, and adjusting for demographic variables (e.g., age) could obscure important real-world difficulties.

This study has several limitations. The population of community-dwelling adults in this study was predominantly Caucasian (95%) and was highly educated. The NOT is a complex task that was developed to be sensitive to the earliest stages of functional impairment in older adults, but it can also be affected by other factors, which must be carefully considered, including English as a second language, learning disabilities, and ADHD. In addition, performance on the NOT can be affected by factors such as motivation and fatigue. Although the props used for the NOT are realistic, the environment is staged and therefore does not provide the same level of support (or distraction) that might be found in the participants' real-world home environment. Participants were also provided with a clearly written list of subtasks that needed to be completed, which is different from many everyday situations. Familiarity with the types of everyday subtasks embedded within the NOT may also have differed across participants. In addition, the sample size for the analysis examining incremental validity based on the in-home performance assessment was small and will require replication. Further work is also needed to assess test–retest reliabilities as most conventional tests of executive functioning (e.g., Wisconsin Card Sorting Test) can only be novel once (Burgess, 1997; Chan et al., 2008).

The NOT represents a promising clinic-based task that may be sensitive to capturing everyday difficulties that can occur in real-world situations, which are often high in cognitive load and ill-structured. The NOT requires simultaneous application, prioritization, and monitoring of goals and sub-goals, inhibition of irrelevant and inappropriate actions to task stimuli, and accurate rule following. The high cognitive load condition required by the NOT may allow for earlier detection of everyday functional difficulties compared with current cognitive tasks. More specifically, in comparison to HOA, we found that individuals with MCI took longer to complete the NOT, were less accurate, less efficient, and committed more inefficient and omission errors. Further, although the MCI group engaged in more planning behaviors and made more self-corrections than the HOA, unlike HOA, these compensatory strategies were not enough to mitigate the poorer performances of the MCI group. The NOT app, which allows for continuous data collection and easier coding of errors and strategies, provides the opportunity to consider not only what the participant can do accurately but also what kind of errors were made and what strategies were used. This information could improve rehabilitation efforts by increasing the precision of training techniques that better target improving functioning in the individual's everyday environment. The NOT showed excellent inter-rater reliability, external validity by distinguishing between age and diagnostic groups, convergent and discriminative validity with standard neuropsychological and other measures, and incremental validity for the prediction of everyday functioning in the aging population. Additional research is needed to better understand the psychometric properties of the NOT primary and process measures and to develop norms.

Acknowledgments

We thank Sarah Norman, Stephanie Saltness and Justin Frow for their assistance in coordinating data collection. We also thank members of the WSU Neuropsychology and Aging laboratory for their help in collecting and scoring the data.

Funding

This study was partially supported by H. L. Eastlick Professorship funding; the National Institute of Biomedical Imaging and Bioengineering (Grant R01 EB009675); and the U.S. Department of Education: Graduate Assistance in Areas of National Need (GAANN funding; Grant P200A150115).

Conflict of Interest

None declared.

References

Albert, M. S., DeKosky, S. T., Dickson, D. et al. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendation from the National Institute of Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(*3*), 270–279.

Alderman, N., Burgess, P. W., Knight, C., & Henman, C. (2003). Ecological validity of a simplified version of the multiple errands shopping test. *Journal of the International Neuropsychological Society*, *9*(*1*), 31–44. doi: 10.1017/s1355617703910046.

Bailey, H. R., Kurby, C. A., Giovannetti, T., & Zacks, J. M. (2013). Action perception predicts action performance. *Neuropsychologia*, *51*(*11*), 2294–2304. doi: 10.1016/j.neuropsychologia.2013.06.022.

Baum, C. M., Morrison, T., Hahn, M., & Edwards, D. F. (2003). *Test manual: Executive function performance test*. St. Louis, MO: Washington University.

Brandt, J., & Folstein, M. (2003). *Telephone interview for cognitive status*. Lutz, FL: Psychological Assessment Resources, Inc.

Burgess, P. W. (1997). Theory and methodology in executive function research. In Rabbit, P. (Ed.), *Methodology of frontal and executive function* (pp. 81–116). East Sussex, United Kingdom: Psychology Press Publishers.

Burgess P. W. (2000). Strategy application disorder: the role of the frontal lobes in human multitasking. *Psychological research*, *63*(*3–4*), 279–288. https://doi.org/10.1007/s004269900006.

Burgess, P. W., Alderman, N., Forbes, C. et al. (2006). The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology. *Journal of the International Neuropsychological Society*, *12*(*2*), 194–209.

Bulzacka, E., Delourme, G., Hutin, V. et al. (2016). Clinical utility of the multiple errands test in schizophrenia: A preliminary assessment. *Psychiatry research*, *240*, 390–397. doi: 10.1016/j.psychres.2016.04.056.

Burns, S. P., Dawson, D. R., Perea, J. D. et al. (2019). Associations between self-generated strategy use and MET-home performance in adults with stroke. *Neuropsychological Rehabilitation*. doi: 10.1080/09602011.2019.1601112.

Burton, C. L., Strauss, E., Bunce, D., Hunter, M. A., & Hultsch, D. F. (2009). Functional abilities in older adults with mild cognitive impairment. *Gerontology*, *55*(*5*), 570–581. doi: 10.1159/000228918.

Castiel, M., Alderman, N., Jenkins, K., Knight, C., & Burgess, P. (2012). Use of the multiple errands test – simplified version in the assessment of suboptimal effort. *Neuropsychological Rehabilitation*, *22*(*5*), 734–751.

Chan, R. C., Shum, D., Toulopoulou, T., & Chen, E. Y. (2008). Assessment of executive functions: Review of instruments and identification of critical issues. *Archives of Clinical Neuropsychology*, *23*(*2*), 201–216. doi: 10.1016/j.acn.2007.08.010.

Cipresso, P., Albani, G., Serino, S. et al. (2014). Virtual Multiple Errands Test (VMET): A virtual reality-based tool to detect early executive functions deficit in Parkinson's disease. *Frontiers in Behavioral Neuroscience*, *8*, 405. doi: 10.3389/fnbeh.2014.00405.

Clark, A. J., Anderson, N. D., Nalder, E., Arshad, S., & Dawson, D. R. (2017). Reliability and construct validity of a revised Baycrest multiple errands test. *Neuropsychological Rehabilitation*, *27*(*5*), 667–684. doi: 10.1080/09602011.2015.1117981.

Craik, F. I., & Bialystok, E. (2006). Planning and task management in older adults: Cooking breakfast. *Memory & Cognition*, *34*(*6*), 1236–1249. doi: 10.3758/bf03193268.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan executive function system: Examiner's manual*. San Antonio, TX: The Psychological Corporation.

Dassel, K. B., & Schmitt, F. A. (2008). The impact of caregiver executive skills on reports of patient functioning. *The Gerontologist*, *48*(*6*), 781–792. doi: 10.1093/geront/48.6.781.

Dawadi, P., Cook, D., & Schmitter-Edgecombe, M. (2013). Automated cognitive health assessment using smart home monitoring of complex tasks. *IEEE Transactions on Systems, Man and Cybernetics*, *43*, 1302–1313. doi: 10.1109/TSMC.2013.2252338.

Dawson, D. R., Anderson, N. D., Burgess, P., Cooper, E., Krpan, K. M., & Stuss, D. T. (2009). Further development of the multiple errands test: Standardized scoring, reliability, and ecological validity for the Baycrest version. *Archives of Physical Medicine and Rehabilitation*, *90*(*11 Suppl*), S41–S51. doi: 10.1016/j.apmr.2009.07.012.

Tomaszewski Farias, S., Gravano, J., Weakley, A. et al. (2020). The everyday compensation (EComp) questionnaire: Construct validity and associations with diagnosis and longitudinal change in cognition and everyday function in older adults. *Journal of the International Neuropsychological Society*, *26*(*3*), 303–313. doi: 10.1017/S135561771900119X.

Feinkohl, I., Cress, U., & Kimmerle, J. (2016). Reheating breakfast: Age and multitasking on a computer-based and a non-computer-based task. *Computers in Human Behavior*, *55*, 432–438.

Finlayson, M., Havens, B., Holm, M. B., & Van Denend, T. (2003). Integrating a performance-based observation measure of functional status into a population-based longitudinal study of aging. *Canadian Journal on Aging/La Revue Canadienne du Vieillissement*, *22*(*2*), 185–195. doi: 10.1017/S0714980800004505.

Giovannetti, T., Bettcher, B. M., Brennan, L., Libon, D. J., Kessler, R. K., & Duey, K. (2008). Coffee with jelly or unbuttered toast: Commissions and omissions are dissociable aspects of everyday action impairment in Alzheimer's disease. *Neuropsychology*, *22*(*2*), 235–245. doi: 10.1037/0894-4105.22.2.235.

Gold, D. A. (2012). An examination of instrumental activities of daily living assessment in older adults and mild cognitive impairment. *Journal of Clinical and Experimental Neuropsychology*, *34*(*1*), 11–34.

Goverover, Y., & Josman, N. (2004). Everyday problem solving among four groups of individuals with cognitive impairments: Examination of the discriminant validity of the observed tasks of daily living—Revised. *OTJR: Occupation, Participation and Health*, *24*, 103–112. doi: 10.1177/153944920402400304.

Jak, A. J., Bondi, M. W., Delano-Wood, L., & Delis, D. C. (2009). Quantification of five neuropsychological approaches to defining mild cognitive impairment. *The American Journal of Psychiatry*, *17*, 368–375.

Kline, P. (2000). *Handbook of psychological testing* (2nd ed.). New York: Routledge.

Knight, C., Alderman, N., & Burgess, P. W. (2002). Development of a simplified version of the Multiple Errands Test for use in hospital settings. *Neuropsychological Rehabilitation*, *12*(*3*), 231–255. doi: 10.1080/09602010244000039.

Kosowicz, M., & MacPherson, S. E. (2017). Improving multitasking assessment in healthy older adults using a prop-based version of the breakfast task. *Applied Neuropsychology Adult*, *24*(*3*), 252–263. doi: 10.1080/23279095.2015.1136310.

Lamberts, K. F., Evans, J. J., & Spikman, J. M. (2010). A real-life, ecologically valid test of executive functioning: The executive secretarial task. *Journal of Clinical and Experimental Neuropsychology*, *32*(*1*), 56–65.

Levine, B., Dawson, D., Boutet, I., Schwartz, M. L., & Stuss, D. T. (2000). Assessment of strategic self-regulation in traumatic brain injury: Its relationship to injury severity and psychosocial outcome. *Neuropsychology*, *14*(*4*), 491–500.

Loeb, P. A. (1996). *Independent Living Scales (ILS) stimulus booklet*. San Antonio, TX: Psychological Corporation.

Manly, T., Hawkins, K., Evans, J., Woldt, K., & Robertson, I. H. (2002). Rehabilitation of executive function: facilitation of effective goal management on complex tasks using periodic auditory alerts. *Neuropsychologia*, *40*(*3*), 271–281. https://doi.org/10.1016/s0028-3932(01)00094-x.

Marson, D., & Hebert, K. R. (2006). Functional Assessment. In Attix, D. K., & Welsh-Bohmer, K. A. (Eds.), *Geriatric neuropsychology: Assessment and intervention* (pp. 158–197). New York, NY, US: Guilford Publications.

Martyr, A., & Clare, L. (2012). Executive function and activities of daily living in Alzheimer's disease: A correlational meta-analysis. *Dementia and Geriatric Cognitive Disorders*, *33*(*2–3*), 189–203. doi: 10.1159/000338233.

McAlister, C., & Schmitter-Edgecombe, M. (2013). Naturalistic assessment of executive function and everyday multitasking in healthy older adults. *Aging, Neuropsychology and Cognition*, *20*, 735–756.

McAlister, C., Schmitter-Edgecombe, M., & Lamb, R. (2016). Examination of variables that may affect the relationship between cognition and functional status in individuals with mild cognitive impairment: A meta-analysis. *Archives of Clinical Neuropsychology*, *31*, 123–147.

Meijer, A. M., & Krampe, R. T. (2018). Movement timing and cognitive control: Adult-age differences in multi-tasking. *Psychological Research*, *82*(*1*), 203–214. doi: 10.1007/s00426-017-0876-4.

Parsons, T. D., & Barnett, M. (2017). Validity of a newly developed measure of memory: Feasibility study of the virtual environment grocery store. *Journal of Alzheimer's Disease*, *59*(*4*), 1227–1235. doi: 10.3233/JAD-170295.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Anger Cella, D. (2011). Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS): Depression, anxiety, and Anger. *Assessment*, *18*(*3*), 263–283.

Podsiadlo, D., & Richardson, S. (1991). The timed "up & go": A test of basic functional mobility for frail elderly persons. *Journal of the American Geriatrics Society*, *39*(*2*), 142–148. doi: 10.1111/j.1532-5415.1991.tb01616.x.

Rand, D., Basha-Abu Rukan, S., Weiss, P. L., & Katz, N. (2009). Validation of the virtual MET as an assessment tool for executive functions. *Neuropsychological Rehabilitation*, *19*(*4*), 583–602. doi: 10.1080/09602010802469074.

Reitan, R. M., & Wolfson, D. (1985). *The Halstead-Reitan Neuropsychological Test Battery: Therapy and clinical interpretation*. Tucson, AZ: Neuropsychological Press.

Rey, A. (1964). Auditory verbal learning test. In *Psychological appraisal of children with cerebral deficits*. Cambridge, Massachusetts: I96 IHarvard University Press.

Robertson, K., & Schmitter-Edgecombe, M. (2017). Naturalistic tasks performed in realistic environments: A review with implications for neuropsychological assessment. *The Clinical Neuropsychologist*, *31*(*1*), 16–42. doi: 10.1080/13854046.2016.1208847.

Rocke, K., Hays, P., Edwards, D., & Berg, C. (2008). Development of a performance assessment of executive function: The Children's kitchen task assessment. *The American Journal of Occupational Therapy*, *62*(*5*), 528–537.

Royall, D. R., Lauterbach, E. C., Kaufer, D., Malloy, P., Coburn, K. L., & Black, K. J. (2007). The cognitive correlates of functional status: A review from the committee on research of the American neuropsychiatric association. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *19*, 249–265.

Sanders, C., Low, C., & Schmitter-Edgecombe, M. (2014). Assessment of planning abilities in individuals with mild cognitive impairment using and open-ended problem-solving task. *Journal of Clinical and Experimental Neuropsychology*, *36*(*10*), 1084–1097. doi: 10.1080/13803395.2014.983462.

Sanders, C., & Schmitter-Edgecombe, M. (2012). Identifying the nature of impairment in planning ability with normal aging. *Journal of Clinical and Experimental Neuropsychology*, *34*, 724–737. doi: 10.1080/13803395.2012.670210.

Sbordone, R. J. (1996). Ecological validity: some critical issues for the neuropsychologist. In Sbordone, R. J., & Long, C. J. (Eds.), *Ecological Validity of Neuropsychological Testing* (pp. 15–41). Delray Beach, FL: GR Press/St. Lucie. Press.

Seidel, G. A., Giovannetti, T., Price, C. C. et al. (2013). Neuroimaging correlates of everyday action in dementia. *Journal of Clinical and Experimental Neuropsychology*, *35*(*9*), 993–1005. doi: 10.1080/13803395.2013.844773.

Schmitter-Edgecombe, M., & Cunningham, R. (2019). *Night Out Task and tablet application*. Clinician Manual.

Schmitter-Edgecombe, M., McAlister, C., & Weakley, A. (2012). Naturalistic assessment of everyday functioning in individuals with mild cognitive impairment: The day out task. *Neuropsychology*, *26*(*5*), 631–641.

Schmitter-Edgecombe, M., Parsey, C., & Cook, D. J. (2011). Cognitive correlates of functional performance in older adults: Comparison of self-report, direct observation, and performance-based measures. *Journal of the International Neuropsychological Society*, *17*(*5*), 853–864. doi: 10.1017/S1355617711000865.

Schmitter-Edgecombe, M., & Parsey, C. M. (2014). Assessment of functional change and cognitive correlates in the progression from healthy cognitive aging to dementia. *Neuropsychology*, *28*(*6*), 881–893. https://doi.org/10.1037/neu0000109.

Schmitter-Edgecombe, M., Parsey, C., & Lamb, R. (2014). Development and psychometric properties of the independent activities of daily living: Compensation scale. *Archives of Clinical Neuropsychology*, *29*(*8*), 776–792.

Schmitter-Edgecombe, M., Woo, E., & Greeley, D. (2009). Characterizing multiple memory deficits and their relation to everyday functioning in individuals with mild cognitive impairment. *Neuropsychology*, *23*, 168–177. doi: 10.1037/a0014186.

Schwartz, M. F., Buxbaum, L. J., Ferraro, M., Veramonti, T., & Segal, M. (2003). *Naturalistic action test*. Suffolk, England: Pearson Assessment.

Shallice, T., & Burgess, P. W. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain*, *114*, 727–741.

Turner, G. R., Novakovic-Agopian, T., Kornblith, E., Adnan, A., Madore, M., Chen, A. J. W. et al. (2020). Goal-oriented attention self-regulation (GOALS) training in older adults. *Aging & Mental Health*, *24*(*3*), 464–473. doi: 10.1080/13607863.2018.1534080.

Weakley, A., Weakley, A. T., & Schmitter-Edgecombe, M. (2019). Compensatory strategy use improves real-world functional performance in community dwelling older adults. *Neuropsychology*, *33*(*8*), 1121–1135. doi: 10.1037/neu0000591.

Wechsler, D., & Psychological Corporation, & Psych Corp (Firm) (2008). *WAIS-IV technical and interpretive manual*. San Antonio, TX: Pearson.

Wilson, B. A., Alderman, N., Burgess, P. W., Emslie, H., & Evans, J. J. (1996). *Behavioural assessment of the dysexecutive syndrome*. Bury St Edmunds: Thames Valley Test Company.

Yu, L., Buysse, D. J., Germain, A. et al. (2011). Development of short forms for PROMIS sleep disturbance and sleep-related impairment item banks. *Behavioral Sleep Medicine*, *10*(*1*), 6–24.

Appendix A

Night out task instructions

For this next task, I want you to imagine that you are in a home environment. Please note the locations of Cup boards A and B. Also note the location of the organizer, hot water pitcher and travel bag (examiner points item locations out to participant as talking). I am now going to have you complete a task that an individual might complete in their own home. Rather than pretending, please actually complete all aspects of the task.

You will also be using this kitchen timer for the task. Please note that you only need to press this button (show participant the big button) once to start the timer.

I am now going to have you complete a variety of tasks that an individual might complete when preparing for a night out. Your evening out will include meeting a friend at the movie theater to see the last showing of the movie, *Muppets Most Wanted*, and later traveling to their house for dessert. To prepare for the night out, you will need to figure out what time you need to leave the house to get to the movie theater. The movie schedule is located in the organizer on the table.

This (night-out list) paper contains a list of all the tasks you will be asked to complete. You will need to record on this piece of paper (show participants the paper with the questions) the cost of the movie for you and your similarly-aged friend, the movie start time, and what time you will need to leave your house to get to the movie theater 10 min before the movie starts. It takes approximately 15 min to get to the theater. (Remove night-out list from view so participant does not start planning).

So that the movie is more enjoyable, you will need to make a thermos of tea to take with you. You can find the tea bags and thermos in Cupboard A. Please use the hot water pitcher to fill your thermos with water. The tea bag must sit in the water for 3 min for the tea to be ready so you will need to start the kitchen timer, which is set for 3 min.

You will also need to locate one snack to take to the movies for you and your friend and gather the correct amount of money for the movies. You will find money in the organizer on the table and snacks in Cupboard A. Your friend loves milk chocolate but does not like dark chocolate.

Just before leaving the house, you will need to call your friend and tell them that you are leaving. Your cell phone is located on the table in the organizer. The phone will not actually turn on, but complete the task as if it does.

Because you plan to make a dessert with your friend after the movie, you will need to locate a recipe for Puppy Chow Bars in the recipe booklet. The recipe booklet is located in Cupboard B. You will also need to gather the items required for the recipe. You will find ingredients in Cupboards A and B. You will need to pack all of the collected items into the travel bag, which can be found hanging on the (chair/door handle/cupboard OR over there—and point), and then walk out the door with all of your items. Do you have any questions?

Remember, we want you to complete all tasks as instructed while multitasking and interweaving the tasks in a way that feels natural and most efficient. When you have finished the tasks, and items have been packed into the travel bag, take the travel bag to the front door and then walk out the door. You can begin whenever you are ready (hand participant both the night-out list and a pen).

Appendix B

Subtask Completion Scores Summed to Compute the NOT Accuracy Score

Each of the 8 subtasks (i.e., movie theatre, tea, snack, change, phone call to friend, recipe, travel bag, exit) is assigned one of the following scores and then summed (range = 8 - 32).

1 = *Efficient*. Assigned when a participant completes the activity with no task errors recorded.
2 = *Inefficient*. Assigned when a participant completes the activity with one or more errors that while slowing or otherwise influencing the efficiency of the completion of the task, the task can still be completed. Such as taking too much money, waiting around for the tea, or missing nonessential ingredients from the recipe. The inefficient errors for each task are represented on the tablet by the light gray buttons.
3 = *Incomplete/Inaccurate*. Assigned when a participant completes the activity with one or more errors that will keep the participant from being able to fulfill the task goal, so that the activity will either be left incomplete or completed inaccurately. Examples include: not taking enough money, not putting the tea bag in the water, or missing essential ingredients from recipe. Incomplete and inaccurate errors are represented by the dark gray buttons on the tablet.
4 = *Never Attempted*. Assigned when a participant never initiates a particular task.

Each activity completion score is calculated based on the error buttons that were pressed as the participant completed the task. If a task has multiple errors, the highest score error is recorded for that task. For example, if a task had one inefficient error and one incomplete error, the completion score for that task would be a 3, Incomplete.

Task Sequences Summed to Compute the NOT Efficiency Score

Total number of the six activities below correctly sequenced (range = 0 – 6).

1. Tea started as one of first four activities.
2. Travel bag retrieved as one of first four activities.
3. Cost of movie determined prior to first attempt at retrieving change.
4. Recipe read prior to retrieving food items.
5. Phone call placed near end (no new task aside exit initiated).
6. Travel bag moved to front door as one of last two activities (exit).