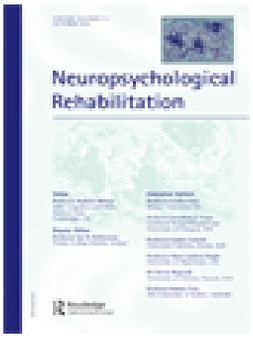

Routledge
Taylor & Francis Group

Neuropsychological Rehabilitation

An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/pnrh20

The Oxford digital multiple errands test (OxMET): Validation of a simplified computer tablet based multiple errands test

Sam S. Webb , Anders Jespersen , Evangeline G. Chiu , Francesca Payne , Romina Basting , Mihaela D. Duta & Nele Demeyere

To cite this article: Sam S. Webb , Anders Jespersen , Evangeline G. Chiu , Francesca Payne , Romina Basting , Mihaela D. Duta & Nele Demeyere (2021): The Oxford digital multiple errands test (OxMET): Validation of a simplified computer tablet based multiple errands test, Neuropsychological Rehabilitation, DOI: 10.1080/09602011.2020.1862679

To link to this article: https://doi.org/10.1080/09602011.2020.1862679

Published online: 06 Jan 2021.

NEUROPSYCHOLOGICAL REHABILITATION
https://doi.org/10.1080/09602011.2020.1862679

Routledge
Taylor & Francis Group

The Oxford digital multiple errands test (OxMET): Validation of a simplified computer tablet based multiple errands test

Sam S. Webb, Anders Jespersen, Evangeline G. Chiu, Francesca Payne, Romina Basting, Mihaela D. Duta and Nele Demeyere

Department of Experimental Psychology, University of Oxford, Oxford, UK

ABSTRACT

Impairments in executive functioning are common following Acquired Brain Injury, though there are few screening tools which present a time efficient and ecologically valid approach to assessing the consequences of executive impairments. We present the Oxford Digital Multiple Errands Test (OxMET), a novel and simplified computer-tablet version of a Multiple Errands Test. We recruited 124 neurologically healthy controls and 105 stroke survivors to complete the OxMET task. Normative data and internal consistency were established from the healthy control data. Convergent and divergent validation was assessed in a mixed subset of 158 participants who completed the OxMET and OCS-Plus. Test–retest reliability was examined across a mixed subset of 39 participants. Finally, we investigated the known-group discriminability of the OxMET. The OxMET demonstrated very high internal consistency, and stable group level test–retest performance as well as good convergent and divergent validity. The OxMET demonstrated high sensitivity and good specificity in overall differentiation of stroke survivors from controls. The Oxford Digital Multiple Errands Test is a brief, easy to administer tool, designed to quickly screen for potential consequences of executive impairments in a virtual environment shopping task on a computer tablet. Initial normative data and validation within a chronic stroke cohort is presented.

ARTICLE HISTORY
Received 24 July 2020
Accepted 8 December 2020

KEYWORDS
Executive function; Stroke; Cognitive impairment; Multiple errands test; Computer tablet

Introduction

Cognitive impairment in executive function is common after acquired brain injury including stroke (Jokinen et al., 2015; Merriman et al., 2019; Millis et al., 2001). Executive function refers to higher-order cognitive abilities such as planning, shifting tasks, and inhibiting behaviour in order to adapt to novel

CONTACT Nele Demeyere nele.demeyere@psy.ox.ac.uk Department of Experimental Psychology, University of Oxford, Anna Watts Building, The Radcliffe Observatory Quarter, Woodstock Road, Oxford OX1 2JD, UK
Supplemental data for this article can be accessed https://doi.org/10.1080/09602011.2020.1862679

situations in everyday life (Gilbert & Burgess, 2008). Impairments in executive functioning have been shown to lead to worse functional outcomes including impairments in instrumental activities of daily living (e.g., Connor & Maeir, 2011; Goverover & Josman, 2004; Mole & Demeyere, 2020; Pohjasvaara et al., 2002). Josman and colleagues suggested that accurate examination of executive functioning following brain injury can reduce the burden on costly and hard to access social and health services, through early signposting and support (Josman et al., 2014).

Executive functioning is a notoriously hard to define cognitive phenomenon (Goldstein et al., 2014), with many theories and models attempting to define what constitutes executive function and how this is linked to the frontal lobes (Gilbert & Burgess, 2008; Luria et al., 1966). One established model is the Supervisory Attentional System (SAS) developed by Norman and Shallice (Norman & Shallice, 1980). In brief, the theory posits that everyday human behaviour is automatic and efficient, except where novelty and difficulty are encountered and behavioural schema are to be updated (Norman & Shallice, 1980). In this case the proposed contention scheduling mechanism chooses a new course of action, and an overarching supervisory control system biases the choice where planning is required (Norman & Shallice, 1980; Van der Linden & Andres, 2001).

Many well used tests have been developed on the basis of the SAS theory, including the Tower of London (Shallice, 1982), and The Hayling and Brixton tests (Bielak et al., 2006; Burgess & Shallice, 1997). The Tower of London assesses planning and problem-solving ability, the Hayling test assesses proponent response inhibition, response initiation, and strategy use (Robinson et al., 2015), and the Brixton task assesses updating responses and abstraction of rules (Van Den Berg et al., 2009; Van der Linden & Andres, 2001). Meta-analysis of neuroimaging tasks have found frontal lobe involvement in performance on tasks associated with the SAS (Cieslik et al., 2015), although frontal lobe damage does not always lead to failure on SAS tasks (see Vordenberg et al., 2014). Though executive functions, defined from any theory, are now most often framed to be supported by a diffuse network of white and grey matter (e.g., Sasson et al., 2013), this network is thought to be mediated by the frontal lobes (Antoniak et al., 2019; Bettcher et al., 2016).

One important issue with executive function, and the neuropsychological tests designed to assess this, is the frontal lobe paradox (George & Gilbert, 2018). This paradox arises where an individual with frontal brain damage performs well on tests of executive function yet has profound executive impairments in everyday life (Shallice & Burgess, 1991). Shallice and Burgess reported the case of three patients with neurological damage specific to the frontal cortex, who each sustained frontal brain injuries yet had high intelligence quotients and performed well on cognitive tests including tests of executive function and language (Shallice & Burgess, 1991). Shallice and Burgess

developed an ecological task to bridge the gap between neuropsychological testing and activities of everyday life (Steverson et al., 2017), where three patients were taken to a shopping centre and given multiple tasks to complete. There were rules to follow, as well as a set limit on money to spend and time to take to complete the task. The goal was for the participants to complete the errands efficiently before reporting back to the researcher. This assessed the patient's problem-solving, planning, and monitoring abilities (Antoniak et al., 2019; Shallice & Burgess, 1991). The patients presented different types of errors, in terms of rule breaks, inefficiencies, interpretation failures, and task failures. Further, when compared to nine IQ and age matched controls on the task, each patient performed at less than the typical 5% control performance rate, which suggested the test was effective at detecting impairment (Shallice & Burgess, 1991). This test was subsequently known as the Multiple Errands Test (MET), characterized as a naturalistic and ecologically valid assessment of executive function. It has been suggested since that the MET may not directly measure executive function, and instead measures the effects of executive dysfunction (Antoniak et al., 2019). Alternative views however suggest the MET in fact assesses executive function to a greater extent than traditional abstract executive assessments, on which dysexecutive patients perform well in presence of daily dysfunction (Shallice & Burgess, 1991). On balance, we suggest the MET can be thought of as a test of executive functioning related to functional outcomes and activities of daily life. As an overall broad-spectrum assessment, it is indeed not a highly controlled domain-specific test but has huge potential as a screening test. The purpose of a screening test is to allow us to determine with high probability that a problem is present, and further assessment is required to understand the nature of the problem and the constituting impairments (Roebuck-Spencer et al., 2017).

In a clinical context however, a MET is often not feasible. Most prominently, the need to take patients out into the real world raises practical concerns regarding transport, staff time, and patient safety. Several versions of the MET have since been created (see review Rotenberg et al., 2020) to work around some of these issues. Real-world versions have been adapted for patients in hospitals (Dawson et al., 2009; Knight et al., 2002), shopping centres (Alderman et al., 2003) and large stores (Antoniak et al., 2019), as well as a home-based version (Burns et al., 2018), addressing some of the barriers. However, patients with neurological conditions often have co-occurring motor impairments restricting their ability to complete the task. Virtual reality/computerized versions may be able to provide a solution here, and several have been developed (e.g., Cipresso et al., 2014; Jovanovski et al., 2012; Rand et al., 2005; Raspelli et al., 2012), though these often bring high costs and needs for technical equipment and expertise which are often not readily available. In a recent systematic review, 33 articles reporting a version of the MET were found and their psychometric properties were

assessed. The MET was commonly scored by accuracy of task completion, task omissions and partial omissions, as well as scores regarding rule breaks (Rotenberg et al., 2020), with partial omissions being most sensitive to impairment (Dawson et al., 2009). Furthermore, this review found that for many versions of the MET, there was generally good internal consistency, good inter-rater reliability, sufficient test–retest reliability, good-adequate convergent validity, and good ability to differentiate clinical groups.

With regards to convergent validity, the MET has been found to converge with a variety of standardized neuropsychological measures of executive functioning. For example, the performance on the MET was found to correlate with the Trail Making Test Trail A time and Trail B, and Trail B/A performance (Alderman et al., 2003; Jovanovski et al., 2012; La Paglia et al., 2014), and with inhibition impairment (Burgess et al., 1998). A 2014 systematic review summarized nine papers that included the MET, validated measures of executive function, and participants with acquired brain injury (Quinn, 2014). The review showed that each paper used a diverse range of neuropsychological tests, commonly including subsets of the Behavioural Assessment of Dysexecutive Syndrome (BADS; Wilson et al., 1996), digit span tests, fluency tests, figure drawing tests, story recall tasks, attention tests, and activities of daily living assessments. Convergent associations between the performance on the MET and the following were found: the Modified Six Elements Test (Jovanovski et al., 2012) and Zoo Map test (Rand et al., 2009) of the BADS (although note no convergence found in Erez et al., 2013 or Okahashi et al., 2013), Rivermead Behavioural Memory Test (Wilson et al., 1999), Comprehensive Assessment of Prospective Memory (Waugh, 1999), and Instrumental Activities of Daily Living Scale (Lawton & Brody, 1969).

Evidence for divergent validity would result from comparing the MET to non-executive measures (Rotenberg et al., 2020), such as memory and intelligence (Hanberg et al., 2018). Explicitly testing for divergent validity has however not been common practice in studies on the Multiple Errands Task. The systematic review by Rotenberg et al. (2020) highlighted that no examinations of divergent validity were conducted in the studies included.

Though broadly speaking convergence with executive tasks and divergence from non-executive tasks seems to classify validation in most cases, the heterogeneity of versions of multiple errands tests and the wide variation in the scoring methods makes a direct comparison of psychometric properties difficult (Rotenberg et al., 2020).

So far, previous versions of the MET have failed to consider it within the framework of a feasible short, ecologically valid, screening tool for the executive aspects of activities of daily life inclusive for the clinical reality after acquired brain injury, which will include individuals with pre-existing dementia, as well as with mobility and upper limb impairments. Specifically, in acute brain injury and in-patient neurorehabilitation settings, cognitive screening must be

time efficient and easy to administer to prevent failure to assess cognition appropriately (Demeyere et al., 2015). Up to now, virtual versions of the MET have used joysticks (see, for example, Titov & Knight, 2005), virtual reality wands (Kourtesis et al., 2020a), and desktop keyboard set ups (see, for example, Law et al., 2006), but are often too long to complete and require complex set ups. These can be expensive, even if the cost of virtual reality technology has decreased in recent years (Kourtesis et al., 2020b).

Outside of the Multiple Errands Tests, other short versions of ecologically valid tests of executive function have been developed, including for instance the 15 min Hotel task (Manly et al., 2002), which is similar to a 6 elements task (Shallice & Burgess, 1991). This example is a table top task, which requires a set-up of complex materials and a skilled examiner for administration and scoring. The computer-tablet is an alternative format not yet tried with the MET that is able to address these issues. A computer tablet app version of the MET could provide guided administration, remove the complexity of any material set up, automatic scoring and shorten the time necessary to test, fitting with the increasing tendency to improve cost-effectiveness with computer-tablet technology in our healthcare settings (Bauer et al., 2012; Koski et al., 2011; Pew Research Centre, 2019).

Older adults have become comfortable performing tasks on computer tablets, due to wider adoption of the computer-tablet format (Anderson & Perrin, 2017). A computer-tablet based version of the MET would shorten testing time, making it more appropriate on time pressed clinical settings, and allow assessment of otherwise difficult to assess patients (e.g., those with mobility and upper limb weakness).

Here, we present a new version of the Multiple Errands Test called the Oxford Digital Multiple Errands Test (OxMET) which is performed on a simple computer tablet app interface with stylus pen and takes less than 10 min to conduct, with the majority of controls completing the task within 3 min. This test hopes to improve usability and feasibility of examining impairments in healthcare settings due to the easy administration of the test. Ultimately, we developed the OxMET to serve as a brief screening tool for impairments of executive dysfunction that may impact activities of everyday life. We established the normative performance scores and clinical cut offs, and examined internal reliability, test–retest reliability, convergent and divergent validity, known-group discriminability, and the sensitivity to impairment of the OxMET in a mixed healthy control and unselected stroke cohort.

Methods

We established the normative data for our test using an English speaking neurologically healthy cohort and assessed the validity of the OxMET outcome measures for an unselected stroke survivor cohort between 2014 and 2019.

We examined the measurement properties of the Oxford Digital Multiple Errands Test (OxMET) and established internal consistency, test–retest reliability, and initial convergent and divergent validity with executive function measures in the OCS-Plus (Demeyere et al., 2020). Approval for the study was gained from the Medical Sciences InterDivisional Research Ethics Committee (R51993/RE001) and the National Research Ethics Committee South Central – Oxford C Research Ethics Committee (REC reference: 18/SC/0044, IRAS project ID: 241571). Data are stored on the Open Science Framework (doi 10.17605/OSF.IO/8SUT), due to copyright, the tests used are not open access.

Participants

A convenience sample of 124 healthy controls with no self-reported neurological history and 105 chronic stroke survivors were recruited from established participant databases from the Translational Neuropsychology Research group at the University of Oxford. All 124 controls completed the OxMET to establish normative data. All stroke survivors had a confirmed diagnosis of stroke or TIA and completed the OxMET. Lesion information was taken from clinical notes and confirmed by visual inspection of clinical brain scans. No selection criteria regarding behaviour or lesion location or size were used. The only two exclusion criteria were an inability to stay alert for the duration of testing and incapacity to provide informed consent.

A mixed subset of 158 participants, consisting of 78 controls and 80 stroke survivors completed the validation tests alongside OxMET. Finally, 39 participants (11 controls and 28 stroke survivors) were retested on OxMET to provide test–retest reliability. The initial start of the project only gathered OxMET data for norming and feasibility, with convergent and divergent validation starting in a later phase. Two measures were added later into data collection as part of a further validation project, and as such only 76–79 participants took part in the Zoo Map test from the Behavioural Assessment of Dysexecutive Syndrome (BADS; Wilson et al., 1996), and the Pill Box (Zartman et al., 2013) tasks. The participants who were retested were selected by opportunity sampling when these participants took part in other studies for the lab where there was additional time availability (Table 1).

Materials

The Oxford digital multiple errands test (OxMET)

The OxMET computer tablet shopping task requires participants to buy six items and to answer two questions. Participants are allowed to complete the tasks in any order. Following an explanation of the tablet use and practice using the pen, the participants are provided with standardized instructions:

Table 1. Characteristics of the cohort samples split by control group and stroke survivor group.

Characteristic	All controls ($n = 124$)	All stroke survivors ($n = 105$)	Validation group ($n = 158$, 80 stroke)	Retested group ($n = 39$, 29 stroke)
Age (*M*, *SD*, range)	65.77 (11.64, 21–91.03)	73.60 (12.09, 29.87–93.26)	68.30 (13.16, 21–93)	69.79 (11.95, 31–92.28)
Education (*M*, *SD*, range)	15.60 (3.85, 6–26)	12.57 (3.33, 8–23)	14.18 (3.86, 6–26)	13.37 (3.37, 9–22)
Handedness (R: L)	105:14	85:10	143:14	36:02
Sex (M: F)	56:68	62:43	83:75	26:13
Stroke side (R: L: B)		44:45:06	32:42:02	09:14:02
Months since stroke (*M*, *SD*, range)		31.53 (23.28, 5.37–144.97)	24.39 (13.34, 5.29–64.01)	
Interval between tests (months)				21.31 (10.05, 2.07–35.18)

Note: For age and education means are presented with *SD* and range in parenthesis. Age is calculated as years of age on day of testing (assessment date – date of birth). Some values are missing with missingness of 9.61% in years of education, 4.80% handedness, 2.86% incomplete stroke side details, and .04% unknown time since stroke for some chronically recruited participants.

> On the following screen you will see a street with shops on it. Your task is to buy six items and to answer two questions. You can enter any shop by tapping on its picture. Once inside a shop you can tap on any item to buy it, the price tag will turn green once selected. Once you know the answer to a question you can tap on the question to answer it. There are some rules to follow. You must take as little time as possible and spend as little money as possible. You can only enter a shop in order to buy an item or to answer a question. You must avoid entering a shop more than once. The errands can be done in any order.

Next, the screen in Figure 1 appears. In order to reduce memory demands for this task, the items on the shopping list can be struck through to keep track and

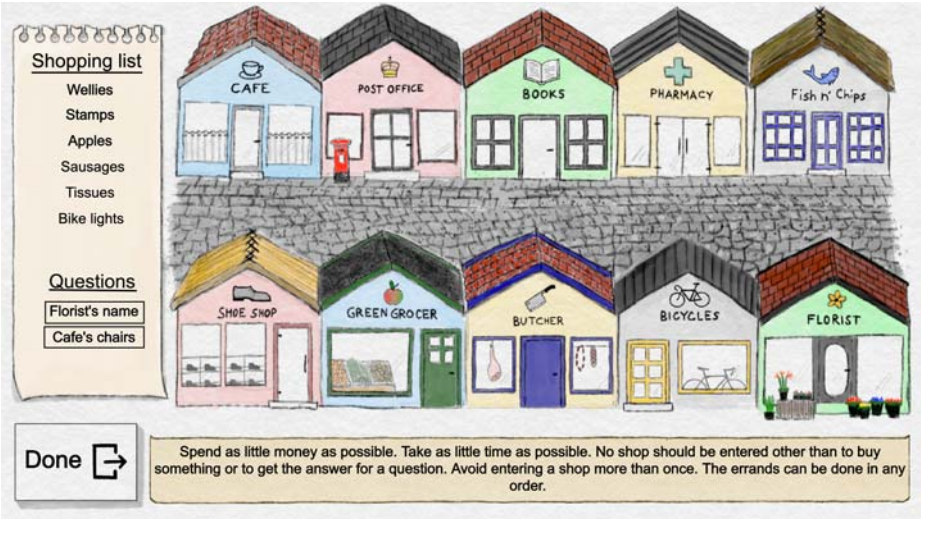

Figure 1. The main street scene from the Oxford Digital Multiple Errands Test where participants first encounter the shopping task. Present are the errands list, questions, and rules. Ten shops are presented, two are distractors and not to be entered at all. Figure available at https://osf.io/rvzmk/ under a CC-BY4.0 license.

the instructions remain on screen at all times. No further elaboration on the instructions is to be given during the task completion in order to standardize administration; Where participants asked questions regarding the task, the researcher would simply point them to the instructions which stayed on the screen at all times. Note, that where there are technical questions from the participants such as how to exit a shop, the examiner can provide help.

The design and "watercolour on paper" look and feel of the task was developed following feedback from stroke survivors where they expressed preference for an adult drawing style over more digitally created elements which were felt to come across as more suitable to children. Participants are able to tap on the image of a shop front with a stylus pen to enter a shop. When inside a shop a static image is presented with six options of items to buy presented on the right of the screen with a price tag (see Figure 2 for an example).

In every shop there are two options of the same type of item at different prices. For instance, where the participant is required to buy apples, there are red and green apples, among the distractor items, and one type is more expensive than the other. Participants are able to tap on the items they desire, the price tag of which turns green when selected, and then pay to leave the store. Alternatively, the participant can buy more than one item or deselect their desired item before paying and leaving. There is no requirement to buy an item in order to leave the store. There is no indication of money spent in total at any time; the participants are implicitly expected to only buy the cheaper option of the item on their shopping list. Ten shops are presented, the "fish and chips" shop and the "books" shop are distractors and not to be entered

Figure 2. The inside of a shop from the Oxford Digital Multiple Errands Test. Participants can select/deselect any item at the right of the shop to turn the price tag green in order to buy it. Figure available at https://osf.io/6h3zf/ under a CC-BY4.0 license.

during the task. A full run-through video of the task is available on the Open Science Framework (doi 10.17605/OSF.IO/8SUT).

The OxMET was run through an application created in MATLAB 2014b on a Microsoft Surface Pro computer tablet (Windows 10 Pro, version 1511) in landscape orientation. The application can run on any windows computer-tablet with touch screen, specific tablet requirements can be found in the supplementary materials (doi 10.17605/OSF.IO/8SUT).

Task scoring

The scoring of the task is completed automatically: no assessor input is required to either save or score the main outcome data. The main outcome measure of the task is accuracy which ranges from minus 10–10 based on a score obtained in each shop. For each of the target buying shops: if the correct shop was entered only once, and when inside only the correct item was bought, then this scored an accuracy point (+1). If the correct shop was not entered, or entered more than once, or entered and a distractor item was bought or no item bought at all, then the participant scores a minus one. For the two task-related shops, an accuracy point was scored if it was entered only once and was left without buying an item and the task question was correctly answered. For the distractor shops, a point was scored if it was not entered.

In sum, if a single error is made in any shop, or a distractor shop is entered, or a question not answered, this deducts a point from the shop related to the task. For instance, the first question asks what the florist's name is, if the florist's shop is not entered the question cannot be scored correctly even if attempted and so the participant is deducted a point for not entering that shop.

In addition to this overall score, the application also stores the breakdown on each type of error and correct move per shop as well as time stamps for full duration of the task and time stamps for time spent in each shop. We calculated additional error scores for comparison to other work using the MET (see Table 2), as well as the total error, which was the sum of frequency of rule breaks, omissions, and commissions. It is possible, however, to generate further scores if this were desired by individual assessors. No requirements were made regarding the order in which the shops needed to be visited.

Validation tests

Participants in the validation group completed brief measures from the Oxford Cognitive Screen – Plus (OCS-plus; see Demeyere et al., 2020), which examined both executive and non-executive abilities. In addition, a more complex measure of planning and executing a complex set of tasks was taken from the Zoo Map test from the Behavioural Assessment of Dysexecutive Syndrome (BADS; Wilson et al., 1996), and an ecological measure of executive functioning in the Pill Box task (Zartman et al., 2013). To establish convergent validity, we compared the OxMET with the executive tasks from the OCS-Plus and the

Table 2. Overview of scoring metrics and descriptions for the Oxford Digital Multiple Errands Test.

Score type	Description	Score range
Accuracy	Total number of correctly completed tasks, where the presence of a one or more errors deducts a total of one point per shop, regardless of number of errors related to the shop.	−10 to 10
Omissions	Number of times participant failed to buy the correct item	0–6 (max six items which should be purchased)
Partial omissions	Number of times the participant: (a) went into a shop but bought the more expensive related item, (b) whether they went into a question-related shop but did not answer the question, (c) or if they answered the question but did not go into the related shop	0–10 (max four points for question related errors) + (max six points for purchasing related items)
Commissions	Number of times the participant bought an unrelated item within a shop	0–24 (max 24 points as there are four distractors per six shops to enter) if each shop is only entered once, where perseveration is present this value can range to infinity
Perseveration	Number of revisits to the same shop	0-infinity (no capped score)
Total error	Sum of: (a) frequency of rule breaks, (b) omissions, and (c) commissions	0-infinity (no capped score)

Zoo Map and Pill Box tasks. Brief measures of language and memory domains from the OCS-Plus were used to establish divergent validity. Note that we considered association between partial omissions and perseveration and the word memory task of the OCS-Plus to be convergent. This is due to how memory deficits would play a part in both forgetting to fully complete tasks and potentially explain why a participant may repeatedly complete an action. An overview of the sub-tasks included in the validation is given in Table 3. The validation tasks were completed in a session lasting maximum one-hour.

Procedure

Participants completed the OxMET on a Microsoft Surface Pro computer tablet (Windows 10 Pro, version 1511) in landscape orientation. The OCS-plus was run in portrait on a separate application on the same device. All tests, including the subtest from the BADS and the Pill Box Test, were conducted by a trained research assistant. Both the OxMET and OCS-Plus were completed within a one-hour session at the Department of Experimental Psychology, Oxford, or at the participants home if they were a stroke survivor. Participants who completed the validation tests completed an additional one-hour session. The researcher sat next to or opposite to the participant when verbally explaining the instructions and during task completion, similar to the video demonstration.

Analysis

Normative data and impairment thresholds for each OxMET measure were calculated in terms of 5th and 95th centiles based upon the performance of the

Table 3. Task descriptions for tasks used to assess hypothesized convergent and divergent validity from the OCS-Plus screen.

Test battery	Measure	Description	Scoring	Validation
OCS-Plus	Picture naming	Non-executive test of comprehension	Total accuracy out of four	Divergent from all OxMET measures
	Semantics	Non-executive test of semantic understanding	Total accuracy out of four	Divergent from all OxMET measures
	Orientation	Non-executive test of participants orientation to time and space	Total accuracy out of four	Divergent from all OxMET measures
	Word memory	Non-executive test of encoding and delayed word recall	Accuracy of two encoding phases and a delayed recall task out of five	Convergent with partial omissions and perseveration
	The Rule Finding and Switching test	Executive test of updating and rule switching	Total accuracy out of 46 and completion time	Convergent with OxMET measures
	The Trail Making Test	Executive test of set shifting	Executive score = ratio of switching/non-switching trails and completion time, and accuracy and completion time on Trail A and B	Convergent with OxMET measures
	Cancellation and Invisible cancellation	Executive test of selective attention and spatial working memory in the invisible cancellation	Total accuracy on both tasks out of 30. False positives and revisits	Convergent with OxMET measures
BADS	Zoo Map	Test of planning, organization, and executive function	Raw total score	Convergent with OxMET measures
Pill Box Test	Pill Box	Test of organization, planning, and executive function	Total errors	Convergent with OxMET measures

Note: Oxford Cognitive Screen – Plus (OCS-Plus). Oxford Digital Multiple Errands Test (OxMET). Behavioural Assessment of Dysexecutive Syndrome (BADS).

healthy control sample in cohort one. We assessed age and education effects on OxMET measures in the control group to determine whether the normative data should be stratified.

Next, we assessed the psychometric properties of the OxMET. Using the normative and stroke survivor data combined for greater power, we established the internal reliability of the OxMET using a split half method and Cronbach's alpha. Further, we assessed test–retest reliability at the group and individual level using the subset of mixed stroke and control sample. Finally, we examined the associations of sub-measures from the OxMET with validation tasks (see Table 3) to establish convergent and divergent validity.

Finally, we assessed the preliminary ability of performance on the app to differentiate healthy controls from stroke survivors, through group comparisons and ROC analysis.

A priori power analysis was not conducted for inferential tests, we instead established a smallest effect size of interest for our correlations ($r = .31$, alpha adjustment described later, 80% power) and calculated power for the one-sided Wilcoxon signed rank tests following Shieh et al. (2006), revealing a power of one with our Bonferroni corrected alpha levels per analyses.

All analyses were computed in R (version 3.5.1; 2018-07-02; R Core Team, 2018), the data and analyses scripts used to generate this manuscript are openly available (doi 10.17605/OSF.IO/8SUT). We used the following packages for analysis and visualizing data: *readxl* (Wickham & Bryan, 2019), *pROC* (Robin et al., 2011), *rcompanion* (Mangiafico, 2019), *sjstats* (Lüdecke, 2018), *Hmisc* (Harrell Jr, 2019), *psych* (Revelle, 2018), *irr* (Gamer et al., 2019), *rstatix* (Kassambara, 2020), *cowplot* (Wilke, 2019), *dplyr* (Wickham et al., 2019), and *wmwpow* (Mollan et al., 2020, R package version 0.1.3).

Results

Normative data

Both age and education (with no imputation for missing data) significantly correlated with most OxMET measures (Bonferroni alpha corrected level for 16 comparisons, $p = .003$), except for omissions, frequency of rule breaks, partial omissions, and perseveration. Correlations can be found in Supplemental Table 1. Through the unbiased code function "split" in base R (R Core Team, 2018) we found three age groups of approximately equal sizes which were 21–63.49 (n = 45), 63.50–71.70 (n = 38), and 71.70–91 (n = 41). This method is unbiased in so far that age group divides are not based on any performance data. The age groups, when compared on OxMET measures, performed statistically different to each other, which therefore justified separate impairment cut offs. The OxMET measures where age groups behaved differently were Time ($H(2) = 14$, $p = .001$), Accuracy ($H(2) = 13.03$, $p = .002$), and Commission errors ($H(2) = 12.59$, $p = .002$). Due to the differences in performance on the OxMET measures between age groups we split the normative data into the age categories. Despite an overall correlation, we found no significant differences when splitting the groups based on a variety of different levels of education and therefore do not present differential education level cut-offs.

Normative data for the OxMET is presented in Table 4 using data taken from the 124 neurologically healthy participants. On average, controls were at ceiling in accuracy and made no errors, except in the over 71 group who on average made two non-specific errors. The main differences in age were apparent in the time to complete the task where the older age groups took longer.

Reliability

Internal consistency

A split half reliability estimate with 5,000 bootstrapped random samples was conducted on OxMET accuracy. Cronbach's alpha revealed an average internal consistency statistic of $a = .79$ ($SD = .13$). We further performed internal consistency analyses on raw scores for all OxMET measures (with reverse coding of

Table 4. Normative medians from 124 neurologically healthy controls for the OxMET stratified by age (scores on subtests lower than 5th centile or higher than 95th centile denote an impairment).

	Overall	≤63.49		63.50–71.69		≥71.70	
Measures	*Med*	*Med*	Centile	*Med*	Centile	*Med*	Centile
Completion Time	194.50	170.28	>285.57	199.38	>278.99	225.86	>397.48
Accuracy	10	10	<6	10	<6	8	<0
Omissions	0	0	>0	0	>0	0	>1
Partial Omissions	0	0	>1	0	>1	0	>2
Frequency of rule breaks	0	0	>2	0	>2	0	>2
Perseverations	0	0	>0	0	>0	0	>1
Commissions	0	0	>1	0	>1	0	>3
Total Errors	0	0	>4	0	>4	2	>6

Note: Oxford Digital Multiple Errands Test (OxMET). All values except centiles are medians. Centiles are 5th for accuracy, and 95th centiles for errors and time. Partial omissions centile was rounded to 2, total error was rounded to 6. *Med* refers to median.

accuracy to be consistent with error scoring) and found a standardized Cronbach's alpha of $a = .87$ (average item correlation $r = .45$, Gutman's Lambda 6 of 1). This demonstrates a very high internal consistency of the OxMET.

Test–retest reliability

Test–retest reliability was assessed on the individual and group level. See Table 5 for Wilcoxon signed rank tests comparing test and retest for differences at the group level, interpretations of *p*-values were Bonferroni corrected (significant if below $p = .00625$). Performance across time for each of the OxMET measures is graphically presented in Supplemental Figure 1 and all intraclass correlation coefficients are presented in Supplemental Table 2.

Validity

The results of the convergent and divergent validity analyses are found in Table 6 and interpretations were corrected for multiple comparisons using the *meff* function in R (Derringer, 2018). This alpha-correlation method is specifically

Table 5. Differences in performance for test–retest of OxMET across time at the group level in a mixed healthy control and stroke sample (n = 39), retested interval (M = 21.31) months (SD = 10.05).

Measure	Test median	Retest median	*W*	*p*	Lower CI	Upper CI
Completion Time	131.6	152.61	493	.15	−5.72	56.59
Accuracy	10	8	336	.03	0	3.50
Omissions	0	1	30	.35	0	1
Partial Omissions	1	0	62	.49	−1.50	1
Frequency of rule breaks	0	2	80	.21	−1.50	1
Perseverations	0	0	24.50	.07	0	2
Commissions	0	0	86.50	.11	−2	0
Total Errors	0	0	183	.20	−2.50	.50

Note: Oxford Digital Multiple Errands Test (OxMET). *W* refers to Wilcoxon signed rank test with continuity correction. Data were computed from a mixed sample of 39 (11 controls and 28 unselected stroke survivors). No significant differences were found at Bonferroni correct level ($p = .00625$).

designed for correcting for non-independent test statistics as in correlations (Derringer, 2018), and thus we use this correction for our large correlational analyses where Bonferroni corrections may be inappropriate. This alpha correction $\alpha_{M_{eff}}$ takes into account the effective number of outcomes from both OxMET ($M_{eff,\ p.\ 11} = 6.05519$) and validation measures ($M_{eff_2} = 22.9746729$). The alpha corrected level for interpretation is $\alpha_{M_{eff}} = \frac{\alpha}{M_{eff1} \times M_{eff2}} = .00036$. Following the definition of convergent validity used by Rotenberg et al. (2020), we interpret convergence if correlations are significant above .30.

With regard to the main outcome measure of the OxMET, accuracy on the OxMET correlated convergently with Trail B accuracy and executive score from the Oxford Cognitive Screen – Plus (OCS-Plus; Demeyere et al., 2020) battery, as well as encoding of words, both attentional tests from the OCS-Plus, and the Zoo Map raw score from the Behavioural Assessment of Dysexecutive Syndrome (BADS; Wilson et al., 1996) but not the Rule Finding test from the OCS-Plus. Accuracy was not associated with comprehension, orientation, delayed memory and non-executive tasks, demonstrating divergent validity.

With regards to the different measures from the OxMET, time taken to complete the OxMET was indiscriminate with its relations, relating to many of the OCS-Plus tasks accuracy measures, but crucially the Trail B time is the one of two time measures that the OxMET time related to, as well as Rule Finding time, which discriminates it from the Trail A baseline and other non-executive tasks. Partial omissions were related to measures of working memory in the immediate word encoding and invisible cancellation accuracy. Commissions related to encoding and delayed recall tasks, as well as Trail B accuracy and executive sore, and Cancellation tasks, suggesting a similar relationship as accuracy. The total error score behaved similar to accuracy and commissions as these are interdependent measures. Full task omissions, frequency of rule breaks, and perseveration scores, did not correlate with any OCS-Plus measure, possibly due to lack of variance and small numbers who made errors.

Discriminability of participant groups

Group comparisons

Wilcoxon rank sum tests with Bonferroni corrections for multiple comparisons ($p = .00625$) and continuity correction were carried out between controls and stroke survivors on each of the OxMET outcome measures (see analysis for variance checks in code justifying this choice of statistic). The groups differed on all OxMET measures except for omissions (see Figure 3 and Table 7). Note we also ran ANCOVA with age and education as covariates to examine the influence of age and education on significance of group comparisons, and this revealed group comparisons were still statistically significant in the same direction even when controlling for the covariates.

Table 6. Convergent and divergent correlations between the OxMET and OCS-Plus in a mixed healthy control and stroke sample (N = 158).

Validation measure	*N*	Time	Accuracy	Omissions	Partial Omissions	Freq. Rule Breaks	Perseveration	Commissions	Total Error
Picture naming	158	−.35*	.17	−.14	−.04	−.21	−.18	−.13	−.21
Semantics	158	−.32*	.13	−.05	−.13	−.19	−.24	−.12	−.16
Orientation	158	−.41*	.15	−.06	−.20	−.11	−.09	−.17	−.17
Encoding 1	158	−.34*	.31*	−.22	−.32*	−.22	−.27	−.32*	−.33*
Encoding 2	158	−.32*	.20	−.10	−.31*	−.17	−.19	−.25	−.23
Delayed word recall	158	−.39*	.25	−.10	−.23	−.24	−.32*	−.26	−.30*
Rule Finding Time	157	.40*	−.12	.15	−.02	.07	.09	.12	.15
Rule Finding Accuracy	157	−.23	.27*	0	−.24	−.19	−.07	−.28*	−.29*
Trails Executive Score	157	−.20	.34*	−.18	−.24	−.17	−.11	−.31*	−.33*
Cancellation Accuracy	155	−.36*	.33*	−.16	−.24	−.18	−.17	−.32*	−.33*
False Positives	155	.17	−.29*	−.03	.27*	.12	.15	.32*	.27
Invisible Cancellation Accuracy	155	−.21	.35*	−.12	−.32*	−.20	−.16	−.35*	−.35*
Zoo Map Raw Error Score	79	−.35	.33	−.28	−.15	−.26	−.32	−.30	−.41*
Pill Box Errors	76	.28	−.27	.21	.20	.20	.23	.14	.30

Note: Data are computed from a mixed sample of 78 healthy controls with no neurological history mixed with 80 unselected stroke survivors. Asterisks denote significance at an alpha corrected level (this alpha corrected level for interpretation is $\alpha_{M_{eff}} = \frac{a}{M_{eff1} \times M_{eff2}} = .00036$).

Sensitivity and specificity

We computed a sensitivity analysis of the main OxMET metric, accuracy to differentiate healthy controls from stroke survivors. We found good sensitivity of the OxMET control 5th centile cut off for accuracy at 74.29% and a specificity of 64.52%. We computed a ROC curve analysis and report an area under the curve of 71.94% (see Figure 4). Dawson et al. suggested that partial omissions may be the best measure to differentiate participant groups (Dawson et al., 2009), and therefore we also computed a ROC analysis on partial omissions and found a sensitivity of 52.38%, a specificity of 77.87%, and an area under the curve of 66.54%. We compared the two ROC results using the *roc.test* function (bootstrap test for two correlated ROC curves) in the *pROC* package and found that the measures were not statistically different at differentiating individuals in groups ($D = -2.79$, $p = 1$). ROC curves for all other OxMET metrics can be found in Supplementary Figure 2.

We further computed ROC analyses for the validation measures regarding the ability to distinguish between stroke survivors and controls and found that in terms of area under the curve (AUC), there were two measures with a greater AUC: the OCS-Plus Encoding 1 and Rule Finding tasks. For sensitivity, the only validation measure with a greater sensitivity was the OCS-Plus delayed recall. In comparison to the Zoo map (AUC = 66.19%) and Pill Box Test (AUC = 70.41%), the OxMET seemed to strike the better balance between high

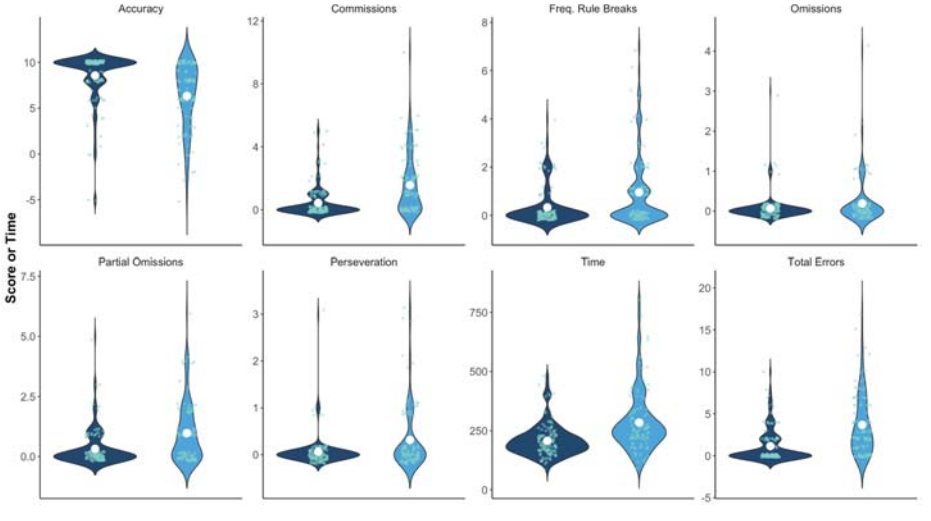

Figure 3. Illustrates performance distributions of healthy controls (dark blue) and stroke survivors (light blue) on all OxMET measures to illustrate differences between groups. White dots represent group-based means on the measure. The stroke survivors largely perform like healthy controls except distinct cases which shift the distribution. Figure available at https://osf.io/8gu6a/ under a CC-BY4.0 license.

Table 7. Performance comparisons between healthy controls and stroke survivors on OxMET scores.

Measure	Control Median	Stroke Median	*W*	*p*	lower CI	upper CI	*R*
Completion Time	194.50	255.72	3586	<.001*	−76.85	−39.41	−.39
Accuracy	10	8	9366.50	<.001*	1	2	.40
Omissions	0	0	5894	.01	0	0	−.16
Partial Omissions	0	1	4286	<.001*	−1	0	−.33
Frequency of rule breaks	0	0	4831	<.001*	0	0	−.28
Perseverations	0	0	5332.50	<.001*	0	0	−.27
Commissions	0	1	3934.50	<.001*	−1	0	−.38
Total Errors	0	3	3252.50	<.001*	−2	−1	−.44

Note: Oxford Digital Multiple Errands Test (OxMET). Multiple comparisons correction was applied ($p = .00625$) to analysis, significance of *p*-values below this level is signified in the table with a single asterisk. Note confidence intervals and *p*-values for all measures except time could not be computed accurately due to ties in the data, thus cautious interpretation is necessary.

sensitivity and specificity. The pillbox test for example demonstrated very high specificity (97.4%) but a low sensitivity (29.74%).

Discussion

We present a standardized new test in the form of a computer tablet app version of the Multiple Errands Test called the Oxford Digital Multiple Errands Test (OxMET). Following the description of the test, psychometric data was

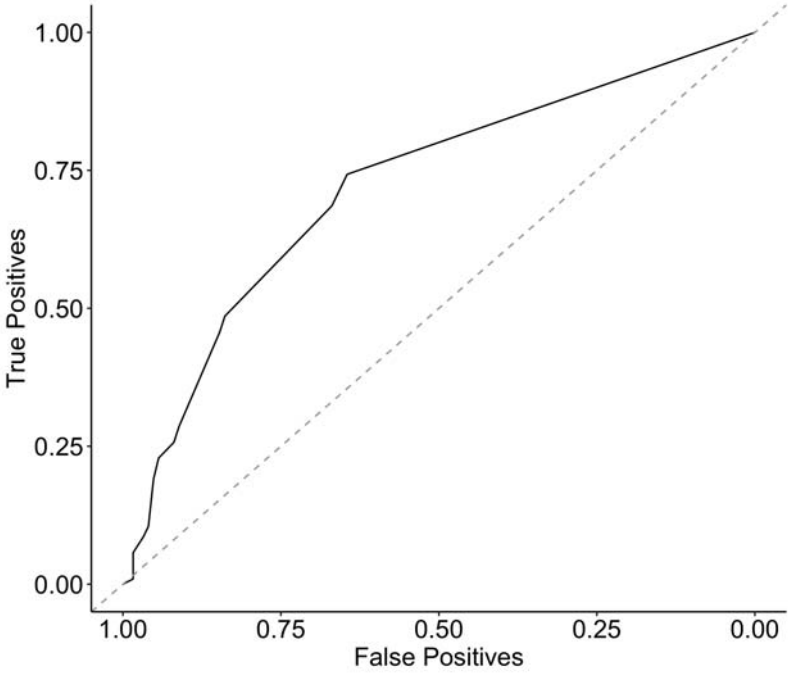

Figure 4. ROC curve of the trade-off between true and false positives in identifying stroke survivors from healthy control centile cut offs for Oxford Digital Multiple Errands Test accuracy. Figure available at https://osf.io/cysg4/ under a CC-BY4.0 license.

provided. We established the association between age and education on the main outcome metric of overall accuracy on the task as well as a range of specific error and time scores. Age based normative cut offs for performance were derived based on a neurologically healthy control cohort of 124 older adults.

The neurologically healthy older adult cohort were found to perform at ceiling on the OxMET metrics, with the exception of older age groups (greater than 71), but no significant differences were found when comparing different education groups on performance. The main difference between groups was the increase in total error score for the older age group, meaning two non-specific errors were made on average in that group, compared to zero in the lower (<63) and middle age groups (63–71). These relative ceiling effects make the OxMET a potentially strong screening measure, with an expected near errorless performance, and any difficulties with this straightforward shopping task likely to demonstrate an impairment in executive functions tapped by the test. Further ecological validity data is required to demonstrate whether this is also likely to flag a significant impact on activities of daily life.

Next, we assessed the reliability of each outcome metric, both internal and across time, and found high internal consistency (trial level $a = .79$; metric level $a = .87$). Furthermore, performance on the OxMET across all metrics demonstrated good test–retest stability on the group level, even with a wide retest interval (average 21.31 months), in a combined test–retest cohort of neurologically healthy adults and stroke survivors.

In addition, we compared performance on the OxMET to a neuropsychological screening tool and two tests of executive function to establish convergent and divergent validity. We found Time for completion to be indiscriminately related to many OCS-Plus Measures and the Zoo Map test, revealing an often-found non-specific performance difference in processing speed and motor slowness in stroke survivors affecting all tasks (e.g., Su et al., 2015). In contrast, the main outcome metric of overall accuracy was more selective in correlations and revealed good convergent and divergent associations above the coefficient of .30 suggested by Rotenberg et al. (2020). This suggests the accuracy measure is a more specific executive and attentional measure, discriminant from lower level basic comprehension and understanding tasks.

The error metrics from the OxMET, which are interdependent on accuracy but differently scored, revealed similar relationships to other variables. The perseveration metric distinctly convergently related to delayed memory recall and the commissions metric related to false positives on an attentional task, suggesting convergence in our measure for items related by inhibition and longer-term memory issues. The frequency with which people made errors did not relate to any comparator measure, suggesting this is not a well understood metric or whether this adds anything to the interpretation of the OxMET over and above specific error types. Our results fit with the mixed results found

by Rotenberg et al. (2020) systematic review, which found inconsistent validity of different outcome measures from multiple errands tasks across 33 studies.

Finally, we compared the healthy control and stroke cohorts on all performance metrics and revealed statistically significant differences (after correction for multiple comparisons) between groups on all OxMET outcome measures bar the omissions measure which did not survive the correction. When further exploring the sensitivity of the OxMETs scores, through data visualization, we found most of the stroke survivor cohort was clustered near the bottom of most error metrics, with a critical number of patients making more errors. With regards to ROC analyses, we found moderate to good sensitivity of the task to differentiating this heterogeneous stroke survivor sample from healthy controls. Though the aim of our test is to screen for executive impairment and not to screen for the presence of a stroke, such pathological group differentiation has been shown in other versions of the MET (Rotenberg et al., 2020, with 12 of the 14 studies examining discriminability showing significant differences). The present results align this digital OxMET version with the known-group literature on multiple errands tasks.

Study limitations and future research

This paper has presented only a subset of the theoretically motivated possible performance metrics that can be derived from the app. For instance, the app stores time stamped information as well as audio-recordings, and other strategies and metrics for completion of the task could be determined and evaluated. We provide all data from this project openly on the Open Science Framework (doi 10.17605/OSF.IO/8SUT5), and would be happy to see further exploration by other researchers.

We did not establish the ecological validity of the OxMET in this investigation, and this is a key next step. In order to firmly establish the link between these virtual and real-world activities of daily life, further validation of OxMET is required. In addition, validating the task as well as linking it to wider functional outcomes is required to establish the informational and clinical value of the OxMET.

Finally, an important issue that affects most Multiple Errands Tests is that they are not easily translatable to different cultures or countries, with obvious examples being hospital specific adaptions (Knight et al., 2002), and ethnocentric requirements in tests such as in Alderman et al. (2003) where they require participants to answer "What is the headline from either today's 'Daily Mail', 'Daily Mirror' or 'The Sun' newspaper?" (p. 44). The OxMET is still biased towards a Western culture, and the design with input from UK stroke survivors intentionally fit a familiar shopping scene. The app has inbuilt technology to be easily translatable to other countries, with flexible pulling in of text files for instructions. With regards to the shop and shopkeeper images, these could

similarly be replaced within the code and cultural adaptations and translations would be encouraged, with new and appropriate norm and acceptability testing required for these versions. There has been a push towards a standardized and adaptable scoring system that can be used across many settings and cultures (see Antoniak et al., 2019 and Burns et al., 2018 for examples), and we feel that although the initial and normed design is UK centric, there is definite scope for adaptation. This approach fits strongly with similar approaches taken by our research group on translations for the Oxford Cognitive Screen (e.g., Robotham et al., 2020; Shendyapina et al., 2019) and OCS-Plus (Demeyere et al., 2020).

Next steps and clinical use

With normative data and initial psychometric investigation of the OxMET now completed, the next steps are to examine the clinical applicability of the app and further validation in specific and selective clinical groups and subgroups. (e.g., in groups with damage to frontal-executive networks such as Shallice & Burgess, 1991), as well as further ecological validation into predictive validity regarding instrumental activities of daily life. The present investigation sets the foundation for future clinical studies to provide the necessary evidence base for clinical adoption.

Once established, the OxMET could fill a clinical need as a brief screen for impairments in executive function that may impact everyday life, which should be further assessed in functional observational assessments of everyday tasks. The brief and inclusive nature of the digital format screen means that all patients can complete this screen to detect potentially hidden impacts of executive impairments on everyday tasks (e.g., see the frontal paradox George & Gilbert, 2018). We believe this screen, if shown to be ecologically valid for real life activities, has the potential to provide key input to decisions around rehabilitation pathways as well as discharge and care package decisions. This may have an especially important contribution when considering a home discharge with independent living. This will however require further research on predictive validity and clinical utility.

Conclusions

The current study presented a novel, tablet based Multiple Errands Task with normative data from a large healthy aging cohort and initial reliability and validation in chronic stroke survivors. We aim for this assessment to provide a quick screening for daily life consequences of executive impairment. Future research should establish further clinical sub-group validation, links to broad functional outcomes, and feasibility of the OxMET assessment in clinical settings.

Disclosure statement

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

Funding

This work was supported by Medical Research Council: [Grant Number MC_PC_14103]; Stroke Association: [Grant Number TSA LECT 2015/02].

Data availability statement

Data deposition: The data and analysis scripts that support the findings of this study are openly available on the Open Science Framework at http://doi.org/10.17605/OSF.IO/8SUT5.

ORCID

Sam S. Webb http://orcid.org/0000-0002-0029-4665
Anders Jespersen http://orcid.org/0000-0001-7607-6536
Evangeline G. Chiu http://orcid.org/0000-0002-8854-5683
Mihaela D. Duta http://orcid.org/0000-0002-0435-571X
Nele Demeyere http://orcid.org/0000-0003-0416-5147

References

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. https://doi.org/10.1017/S1355617703910046

Anderson, M., & Perrin, A. (2017). *Tech adoption climbs among older adults*. Pew Research Center, 1–22.

Antoniak, K., Clores, J., Jensen, D., Nalder, E., Rotenberg, S., & Dawson, D. R. (2019). Developing and validating a big-store multiple errands test. *Frontiers in Psychology*, *10*, 2575. https://doi.org/10.3389/fpsyg.2019.02575

Bauer, R. M., Iverson, G. L., Cernich, A. N., Binder, L. M., Ruff, R. M., & Naugle, R. I. (2012). Computerized neuropsychological assessment devices: Joint position paper of the American academy of clinical neuropsychology and the national academy of neuropsychology. *The Clinical Neuropsychologist*, *26*(2), 177–196. https://doi.org/10.1080/13854046.2012.663001

Bettcher, B. M., Mungas, D., Patel, N., Elofson, J., Dutt, S., Wynn, M., Watson, C. L., Stephens, M., Walsh, C. M., & Kramer, J. H. (2016). Neuroanatomical substrates of executive functions: Beyond prefrontal structures. *Neuropsychologia*, *85*, 100–109. https://doi.org/10.1016/j.neuropsychologia.2016.03.001

Bielak, A. A. M., Mansueti, L., Strauss, E., & Dixon, R. A. (2006). Performance on the hayling and brixton tests in older adults: Norms and correlates. *Archives of Clinical Neuropsychology*, *21*(2), 141–149. https://doi.org/10.1016/j.acn.2005.08.006

Burgess, P. W., Alderman, N., Evans, J., Emslie, H., & Wilson, B. A. (1998). The ecological validity of tests of executive function. *Journal of the International Neuropsychological Society*, *4*(6), 547–558. https://doi.org/10.1017/S1355617798466037

Burgess, P. W., & Shallice, T. (1997). *The Hayling and Brixton tests*. Thames Valley Test Company.

Burns, S. P., Pickens, N. D., Dawson, D. R., Perea, J. D., Vas, A. K., Marquez de la Plata, C., & Neville, M. (2018). In-home contextual reality: A qualitative analysis using the multiple errands test home version (MET-home). *Neuropsychological Rehabilitation*, 1–15. https://doi.org/10.1080/09602011.2018.1431134

Cieslik, E. C., Mueller, V. I., Eickhoff, C. R., Langner, R., & Eickhoff, S. B. (2015). Three key regions for supervisory attentional control: Evidence from neuroimaging meta-analyses. *Neuroscience & Biobehavioral Reviews*, *48*, 22–34. https://doi.org/10.1016/j.neubiorev.2014.11.003

Cipresso, P., Albani, G., Serino, S., Pedroli, E., Pallavicini, F., Mauro, A., & Riva, G. (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. https://doi.org/10.3389/fnbeh.2014.00405

Connor, L. T., & Maeir, A. (2011). Putting executive performance in a theoretical context. *OTJR: Occupation, Participation and Health*, *31*(1_suppl), S3–S7. https://doi.org/10.3928/15394492-20101108-02

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), S41–S51. https://doi.org/10.1016/j.apmr.2009.07.012

Demeyere, N., Haupt, M., Webb, S. S., Strobel, L., Milosevich, E., Moore, M., Wright, H., Finke, K., & Duta, M. (2020, February 25). The Oxford Cognitive Screen – Plus (OCS-Plus): A tablet based short cognitive screening tool for milder cognitive impairment. https://doi.org/10.31234/osf.io/b2vgc

Demeyere, N., Riddoch, M. J., Slavkova, E. D., Bickerton, W. L., & Humphreys, G. W. (2015). The Oxford cognitive screen (OCS): validation of a stroke-specific short cognitive screening tool. *Psychological Assessment*, *27*(3), 883–894. https://doi.org/10.1037/pas0000082

Derringer, J. (2018, April 16). A simple correction for non-independent tests. https://doi.org/10.31234/osf.io/f2tyw

Erez, N., Weiss, P. L., Kizony, R., & Rand, D. (2013). Comparing performance within a virtual supermarket of children with traumatic brain injury to typically developing children: A pilot study. *OTJR: Occupation, Participation and Health*, *33*(*4*), 218–227. https://doi.org/10.3928/15394492-20130912-04

Gamer, M., Lemon, J., & Singh, P. (2019). *irr: Various coefficients of interrater reliability and agreement*. R package version 0.84.1. https://cran.r-project.org/web/packages/irr/index.html

George, M., & Gilbert, S. (2018). Mental capacity act (2005) assessments: Why everyone needs to know about the frontal lobe paradox. In J. Fish (Ed.), *The neuropsychologist* (Vol. 5, pp. 59–66). British Psychological Society.

Gilbert, S. J., & Burgess, P. W. (2008). Executive function. *Current Biology*, *18*(3), R110–R114. https://doi.org/10.1016/j.cub.2007.12.014

Goldstein, S., Naglieri, J. A., Princiotta, D., & Otero, T. M. (2014). *Introduction: A history of executive functioning as a theoretical and clinical construct. Handbook of executive functioning (pp. 3-12)*. Springer.

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*(3), 103–112. https://doi.org/10.1177/153944920402400304

Hanberg, V. L., MacKenzie, D. E., & Merritt, B. K. (2018). Scoping review of the Multiple Errands test: Is it relevant to youths with acquired brain injury? *British Journal of Occupational Therapy*, *81*(12), 673–686. https://doi.org/10.1177/0308022618791714

Harrell Jr, F. E. (2019). *Hmisc: Harrell Miscellaneous*. R package version 4.3-0. https://CRAN.R-project.org/package=Hmisc

Jokinen, H., Melkas, S., Ylikoski, R., Pohjasvaara, T., Kaste, M., Erkinjuntti, T., & Hietanen, M. (2015). Post-stroke cognitive impairment is common even after successful clinical recovery. *European Journal of Neurology*, *22*(9), 1288–1294. https://doi.org/10.1111/ene.12743

Josman, N., Kizony, R., Hof, E., Goldenberg, K., Weiss, P. L., & Klinger, E. (2014). Using the virtual action planning-supermarket for evaluating executive functions in people with stroke. *Journal of Stroke and Cerebrovascular Diseases*, *23*(5), 879–887. https://doi.org/10.1016/j.jstrokecerebrovasdis.2013.07.013

Jovanovski, D., Zakzanis, K., Campbell, Z., Erb, S., & Nussbaum, D. (2012). Development of a novel, ecologically oriented virtual reality measure of executive function: The multitasking in the city test. *Applied Neuropsychology: Adult*, *19*(3), 171–182. https://doi.org/10.1080/09084282.2011.643955

Kassambara, A. (2020). *Rstatix: Pipe-Friendly Framework for basic Statistical tests*. R package version 0.4.0. https://CRAN.R-project.org/package=rstatix

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. https://doi.org/10.1080/09602010244000039

Koski, L., Brouillette, M., Lalonde, R., Hello, B., Wong, E., Tsuchida, A., & Fellows, L. K. (2011). Computerized testing augments pencil-and-paper tasks in measuring HIV-associated mild cognitive impairment. *HIV Medicine*, *12*(8), 472–480. https://doi.org/10.1111/j.1468-1293.2010.00910.x

Kourtesis, P., Collina, S., Doumas, L. A. A., & MacPherson, S. E. (2020a). Validation of the virtual reality everyday assessment Lab (VR-EAL): An Immersive virtual reality neuropsychological battery with Enhanced ecological validity. *Journal of the International Neuropsychological Society*, 1–16. https://doi.org/10.1017/S1355617720000764

Kourtesis, P., Korre, D., Collina, S., Doumas, L. A. A., & MacPherson, S. E. (2020b). Guidelines for the Development of Immersive virtual reality Software for cognitive Neuroscience and Neuropsychology: The Development of virtual reality everyday assessment Lab (VR-EAL), a neuropsychological test battery in Immersive virtual reality. *Frontiers in Computer Science*, *1*. https://doi.org/10.3389/fcomp.2019.00012

La Paglia, F., La Cascia, C., Rizzo, R., Cangialosi, F., Sanna, M., Riva, G., & La Barbera, D. (2014). Cognitive assessment of OCD patients: NeuroVR vs neuropsychological test. *Studies in Health Technology and Informatics*, *199*, 40–44.

Law, A. S., Logie, R. H., & Pearson, D. G. (2006). The impact of secondary tasks on multitasking in a virtual environment. *Acta Psychologica*, *122*(1), 27–44.

Lawton, M. P., & Brody, E. M. (1969). Assessment of older people: Self-maintaining and instrumental activities of daily living. *The Gerontologist*, *9*(3), 179–186.

Lüdecke, D. (2018). *Sjstats: Statistical functions for regression models*. R package version *0.14. 0*.

Luria, A. R., Karpov, B. A., & Yarbuss, A. L. (1966). Disturbances of active visual perception with lesions of the frontal lobes. *Cortex*, *2*(2), 202–212. https://doi.org/10.1016/S0010-9452(66)80003-5

Mangiafico, S. (2019). *Rcompanion: Functions to support Extension education Program Evaluation*. R package version 2.3.7. https://CRAN.R-project.org/package=rcompanion

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

Merriman, N. A., Sexton, E., McCabe, G., Walsh, M. E., Rohde, D., Gorman, A., Jeffares, I., Donnelly, N.-A., Pender, N., Williams, D. J., Horgan, F., Doyle, F., Wren, M.-A., Bennett, K.

E., & Hickey, A. (2019). Addressing cognitive impairment following stroke: Systematic review and meta-analysis of non-randomised controlled studies of psychological interventions. *BMJ Open*, *9*(2), e024429. https://doi.org/10.1136/bmjopen-2018-024429

Millis, S. R., Rosenthal, M., Novack, T. A., Sherer, M., Nick, T. G., Kreutzer, J. S., High, W. M., & Ricker, J. H. (2001). Long-term neuropsychological outcome after traumatic brain injury. *Journal of Head Trauma Rehabilitation*, *16*(4), 343–355. https://doi.org/10.1097/00001199-200108000-00005

Mole, J. A., & Demeyere, N. (2020). The relationship between early post-stroke cognition and longer term activities and participation: A systematic review. *Neuropsychological Rehabilitation*, *30*(*2*), 346–370. https://doi.org/10.1080/09602011.2018.1464934

Mollan, K., Trumble, I., Reifeis, S., Ferrer, O., Bay, C., Baldoni, P., & Hudgens, M. (2020). Precise and accurate power of the rank-sum test for a continuous outcome. *Journal of Biopharmaceutical Statistics*, *30*(4), 639–648. https://doi.org/10.1080/10543406.2020.1730866

Norman, D. A., & Shallice, T. (1980). *Attention to action: Willed and automatic control of behavior technical report no. 8006*.

Okahashi, S., Seki, K., Nagano, A., Luo, Z., Kojima, M., & Futaki, T. (2013). A virtual shopping test for realistic assessment of cognitive function. *Journal of Neuroengineering and Rehabilitation*, *10*(1)), 59. https://doi.org/10.1186/1743-0003-10-59

Pew Research Centre. (2019). *Mobile fact sheet*. http://www.pewinternet.org/fact-sheet/mobile/

Pohjasvaara, T., Leskela, M., Vataja, R., Kalska, H., Ylikoski, R., Hietanen, M., Leppavuori, A., Kaste, M., & Erkinjuntti, T. (2002). Post-stroke depression, executive dysfunction and functional outcome. *European Journal of Neurology*, *9*(3), 269–275. https://doi.org/10.1046/j.1468-1331.2002.00396.x

Quinn, T. (2014). *Development of novel computerised tools to assess memory and planning problems in people with brain injury* (Doctoral dissertation, University of Glasgow). http://theses.gla.ac.uk/5689/1/2014quinndclinpsy.pdf

Rand, D., Katz, N., Shahar, M., Kizony, R., & Weiss, P. L. (2005). The virtual mall: A functional virtual environment for stroke rehabilitation. *Annual Review of Cybertherapy and Telemedicine: A Decade of VR*, *3*, 193–198. https://doi.org/10.1037/e705572011-108

Rand, D., Rukan, S. B. A., Weiss, P. L., & Katz, N. (2009). Validation of the Virtual MET as an assessment tool for executive functions. *Neuropsychological Rehabilitation*, *19*(4), 583–602.

Raspelli, S., Pallavicini, F., Carelli, L., Morganti, F., Pedroli, E., Cipresso, P., Poletti, B., Corra, B., Sangalli, D., Silani, V., & Riva, G. (2012). Validating the neuro VR-based virtual version of the multiple errands test: Preliminary results. *Presence: Teleoperators and Virtual Environments*, *21*(1), 31–42. https://doi.org/10.1162/PRES_a_00077

R Core Team. (2018). *R: A language and environment for statistical computing*.

Revelle, W. (2018). *Psych: Procedures for Personality and Psychological research*. Northwestern University. https://CRAN.R-project.org/package=psych Version = 1.8.12

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 1–8.

Robinson, G. A., Cipolotti, L., Walker, D. G., Biggs, V., Bozzali, M., & Shallice, T. (2015). Verbal suppression and strategy use: A role for the right lateral prefrontal cortex? *Brain*, *138*(4), 1084–1096. https://doi.org/10.1093/brain/awv003

Robotham, R. J., Riis, J. O., & Demeyere, N. (2020). A Danish version of the Oxford cognitive screen: A stroke-specific screening test as an alternative to the MoCA. *Aging, Neuropsychology, and Cognition*, *27*(1), 52–65. https://doi.org/10.1080/13825585.2019.1577352

Roebuck-Spencer, T. M., Glen, T., Puente, A. E., Denney, R. L., Ruff, R. M., Hostetter, G., & Bianchini, K. J. (2017). Cognitive screening tests versus comprehensive neuropsychological test batteries: A national academy of neuropsychology education paper. *Archives of Clinical Neuropsychology*, *32*(4), 491–498. https://doi.org/10.1093/arclin/acx021

Rotenberg, S., Ruthralingam, M., Hnatiw, B., Neufeld, K., Yuzwa, K. E., Arbel, I., & Dawson, D. R. (2020). Measurement properties of the multiple errands test: A systematic review. *Archives of Physical Medicine and Rehabilitation*. https://www.sciencedirect.com/science/article/pii/S0003999320301106

Sasson, E., Doniger, G. M., Pasternak, O., Tarrasch, R., & Assaf, Y. (2013). White matter correlates of cognitive domains in normal aging with diffusion tensor imaging. *Frontiers in Neuroscience*, *7*, 32. https://doi.org/10.3389/fnins.2013.00032

Shallice, I. (1982). *Tower of London test*. Sina Research Institute of Behavioral Cognitive Science (Ravantajhiz), 1387.

Shallice, T., & Burgess, P. W. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain*, *114*(Pt 2), 727–741. https://doi.org/10.1093/brain/114.2.727

Shendyapina, M., Kuzmina, E., Kazymaev, S., Petrova, A., Demeyere, N., & Weekes, B. S. (2019). The Russian version of the Oxford cognitive screen: Validation study on stroke survivors. *Neuropsychology*, *33*(1), 77–92. https://doi.org/10.1037/neu0000491

Shieh, G., Jan, S. L., & Randles, R. H. (2006). On power and sample size determinations for the Wilcoxon–Mann–Whitney test. *Journal of Nonparametric Statistics*, *18*(1), 33–43. https://doi.org/10.1080/10485250500473099

Steverson, T., Adlam, A. R., & Langdon, P. E. (2017). Development and validation of a modified multiple errands test for adults with intellectual disabilities. *Journal of Applied Research in Intellectual Disabilities*, *30*(2), 255–268. https://doi.org/10.1111/jar.12236

Su, C. Y., Wuang, Y. P., Lin, Y. H., & Su, J. H. (2015). The role of processing speed in post-stroke cognitive dysfunction. *Archives of Clinical Neuropsychology*, *30*(2), 148–160. https://doi.org/10.1093/arclin/acu057

Titov, N., & Knight, R. G. (2005). A computer-based procedure for assessing functional cognitive skills in patients with neurological injuries: The virtual street. *Brain Injury*, *19*(5), 315–322.

Van Den Berg, E., Nys, G. M., Brands, A. M., Ruis, C., Van Zandvoort, M. J., & Kessels, R. P. (2009). The brixton spatial anticipation test as a test for executive function: Validity in patient groups and norms for older adults. *Journal of the International Neuropsychological Society*, *15*(5), 695–703. https://doi.org/10.1017/S1355617709990269

Van der Linden, M., & Andres, P. (2001). Supervisory attentional system in patients with focal frontal lesions. *Journal of Clinical and Experimental Neuropsychology*, *23*(2), 225–239. https://doi.org/10.1076/jcen.23.2.225.1212

Vordenberg, J. A., Barrett, J. J., Doninger, N. A., Contardo, C. P., & Ozoude, K. A. (2014). Application of the Brixton spatial anticipation test in stroke: Ecological validity and performance characteristics. *The Clinical Neuropsychologist*, *28*(2), 300–316. https://doi.org/10.1080/13854046.2014.881555

Waugh, N. (1999). *Self-report of the the young, middle-aged, young-old and old-old individuals on prospective memory functioning* (Doctoral dissertation).

Wickham, H., & Bryan, J. (2019). *Readxl: Read Excel files*. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

Wickham, H., François, R., Henry, L., & Müller, K. (2019). *Dplyr: A Grammar of data Manipulation*. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr

Wilke, C. O. (2019). *Cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 1.0.0. https://CRAN.R-project.org/package=cowplot

Wilson, B. A., Clare, L., Cockburn, J., Baddeley, A. D., Tate, R., & Watson, P. (1999). *The Rivermead behavioural memory test-extended version*. Thames Valley Test Corporation.

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

Zartman, A. L., Hilsabeck, R. C., Guarnaccia, C. A., & Houtz, A. (2013). The pillbox test: An ecological measure of executive functioning and estimate of medication management abilities. *Archives of Clinical Neuropsychology*, *28*(4), 307–319. https://doi.org/10.1093/arclin/act014