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

Predictive validity of the Oxford digital multiple errands test (OxMET) for functional outcomes after stroke

Sam S. Webb & Nele Demeyere

To cite this article: Sam S. Webb & Nele Demeyere (2023): Predictive validity of the Oxford digital multiple errands test (OxMET) for functional outcomes after stroke, Neuropsychological Rehabilitation, DOI: 10.1080/09602011.2023.2247152

To link to this article: https://doi.org/10.1080/09602011.2023.2247152

Published online: 17 Aug 2023.

NEUROPSYCHOLOGICAL REHABILITATION
https://doi.org/10.1080/09602011.2023.2247152

CASE REPORT

OPEN ACCESS Check for updates

Predictive validity of the Oxford digital multiple errands test (OxMET) for functional outcomes after stroke

Sam S. Webb [a] and Nele Demeyere [b]

[a]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK; [b]Experimental Psychology, University of Oxford, Oxford, UK

ABSTRACT

The Oxford Digital Multiple Errands Test (OxMET) is a brief computer-tablet based cognitive screen, intended as an ecologically valid assessment of executive dysfunction. We examined aspects of predictive validity in relation to functional outcomes. Participants (≤ 2 months post-stroke) were recruited from an English-speaking stroke rehabilitation in-patient setting. Participants completed OxMET. The Barthel Index, Therapy Outcome Measure (TOMS), and modified Rankin Scale (mRS) were collected from medical notes. Participants were followed up after 6-months and completed the Nottingham Extended Activities of Daily Living (NEADL) scale. 117 participants were recruited ($M = 26.18$ days post-stroke ($SD = 25.16$), mean 74.44yrs ($SD = 12.88$), median NIHSS 8.32 ($IQR = 5\text{-}11$)). Sixty-six completed a follow-up ($M = 73.94$yrs ($SD = 12.68$), median NIHSS 8 ($IQR = 4\text{-}11$)). Significant associations were found between TOMS and mRS. At 6-month follow up, we found a moderate predictive relationship between the OxMET accuracy and NEADL ($R^2 = .29$, $p < .001$), and we did not find this prediction with MoCA taken at 6-months. The subacute OxMET associated with measures of functionality and disability in a rehabilitation context, and in activities of daily living. The OxMET is an assessment of executive function with good predictive validity on clinically relevant functional outcome measures that may be more predictive than other cognitive tests.

ARTICLE HISTORY
Received 16 February 2023
Accepted 2 August 2023

KEYWORDS
Executive function; Stroke; Cognitive impairment; Function; Computer tablet

Introduction

Executive functions are a difficult to measure construct consisting of higher-order cognitive abilities, which relate to planning and organizing of behaviour

CONTACT Sam S. Webb sam.webb@psy.ox.ac.uk Department of Experimental Psychology, Anna Watts building, Radcliffe Observatory Quarter, University of Oxford, Oxford, OX2 6GG, UK

Supplemental data for this article can be accessed online at https://doi.org/10.1080/09602011.2023.2247152.

in response to novel situations or goals (Gilbert & Burgess, 2008). Executive dysfunction is a commonly encountered problem following acquired brain injury, including in stroke (e.g., Merriman et al., 2019) with tangible impacts on activities of daily living (Mole & Demeyere, 2020; Oksala et al., 2009; Pohjasvaara et al., 2002). Difficulty in measuring executive dysfunction can arise where individuals with brain damage perform well on structured tests in a laboratory or clinical setting. But are markedly impaired in their everyday life, sometimes referred to as "the frontal lobe paradox" (Gilbert & Burgess, 2008; Shallice & Burgess, 1991). Structured neuropsychological tests have traditionally had poor predictive or associative relationships with everyday life behaviours (Wilson, 1993).

Widely used tests for executive function, like the trail-making test (Reitan & Wolfson, 1995) (TMT), were not designed to reflect real life and therefore lack verisimilitude. Nevertheless, these tests often show veridicality in their association with activities of daily life (e.g., Lipskaya-Velikovsky et al., 2018). Related to this, and perhaps exemplary to the issue of ecological and face validity, is the frontal lobe paradox, where many individuals still pass these abstract tests while being impaired in everyday life (Gilbert & Burgess, 2008; Shallice & Burgess, 1991). Predictive validity is a form of criterion validity where one evaluates the magnitude of an association between the current test and a criterion test at a later date (American Educational Research Association et al., 2014). For example, Katz, Tadmor, Felzen, and Hartman-Maeir examined the predictive validity of the Behavioural Assessment of Dysexecutive Syndrome (BADS) (Wilson et al., 1997) in predicting functional outcomes for those with schizophrenia, in an approach similar to the current investigation (Katz et al., 2007). Whilst the BADS is one of the most popular cognitive assessments of executive functioning, it takes approximately 40 min to administer, and requires the administrator to have specific qualifications and training (defined at level "A" or "B" qualifications from the Pearson website; https://www.pearsonclinical.co.uk/). Whilst the DEX questionnaire aims to reflect everyday problems and experiences, the battery as a whole does not directly relate performance to a real-life scenario.

The Multiple Errands Test (MET) (Shallice & Burgess, 1991), and adaptions (e.g., Alderman et al., 2003; Antoniak et al., 2019; Burns et al., 2019; Jovanovski et al., 2012; Knight et al., 2002; Morrison et al., 2013; Rand et al., 2009; Webb et al., 2022) attempted to mitigate the discrepancy between performance on structured tests and real life behaviours. The Multiple Errands Test was designed to be ecologically valid by taking participants to an unfamiliar setting to complete multiple semi-structured tasks within rule constraints of time and money limits. The participants are monitored to check their progress with the task and the examiner notes their efficiency and deviations from rules. The MET can reveal deficits in planning, problem solving, and self-monitoring abilities, all abilities under the executive functions umbrella (Antoniak et al., 2019; Shallice & Burgess, 1991).

In a systematic review of the MET literature, Rotenberg et al. (2020) found there were 12 studies that examined the relation of the MET to everyday behaviours with 24 hypotheses from these papers (Rotenberg et al., 2020). There were inconsistent relationships between METs and everyday behaviours, with 11 of 24 hypotheses being confirmed (Rotenberg et al., 2020), and mixed evidence being found even when using the same measure across studies. The most frequently used measure to reflect real life problems was the dysexecutive questionnaire (DEX) taken from the BADS, and many MET studies did not report consistent associations of their METs with the DEX (Rotenberg et al., 2020). However, few MET studies looked at the relation to activities of daily living explicitly through assessments directly assessing abilities, thus whilst the face validity of the MET is extremely good, its predictive validity for real life outcomes remains unclear.

We recently developed a computer-tablet version of the test, named the Oxford Digital Multiple Errands Test (OxMET) to be run at bedside as a screening tool (Webb et al., 2022). This app-based version has core advantages of accessibility and feasibility, by reducing demands on staff time and logistics, with a limited cost of a standard tablet and allowing inclusion of those with mobility and language impairments.

Whilst several other computerized or virtual METs have been developed (Cipresso et al., 2014; Jovanovski et al., 2012; Rand et al., 2005; Raspelli et al., 2012), these often bring high costs associated with equipment and technical expertise, and are not easily available. In addition, several require physically moving a patient to a specialized set-up in order to conduct the test, which is often not practical or possible in hospital settings. The OxMET addresses these shortcomings of other METs by being quick to administer and score and being available in a cross-platform application format through the standard appstores. In comparison to other neuropsychological tests and other METs, the OxMET attempts to reflect a real-life scenario of shopping in a way that covers both verisimilitude and veridicality, but that can be done in less than five minutes, even with severely impaired populations. In addition, the OxMET has a significant evidence base demonstrating its psychometrics, in particular having convergent validity both to abstract executive functioning tests (Webb et al., 2022) as well as to real life behaviours (Webb et al., 2022) demonstrating both verisimilitude and veridicality.

Here we set out to determine the predictive validity of the OxMET with regards to meaningful functional outcomes both in a subacute stroke rehabilitation setting and longer-term in stroke recovery (6-months post-stroke). We investigated the association between the performance of a stroke sample on OxMET and routine tests of functional ability collected from a stroke rehabilitation unit. We further followed up a consecutive proportion of the sample to examine predictive abilities of the OxMET taken sub acutely to activities of daily living 6 months later.

This study was pre-registered, and data and analysis scripts to recreate the manuscript are openly available in CC-BY 4.0 license (doi.org/10.17605/OSF.IO/WEP3V).

Methods

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study, following the approach for transparency in research by Simmons et al. (2012). Our manuscript adheres to the STROBE cohort study check-list (von Elm et al., 2007).

This study protocol was approved by the National Research Ethics Committee South Central – Oxford C Research Ethics Committee (OCS-RECOVERY; REC reference: 18/SC/0044, IRAS project ID: 241571). All participants provided written or witnessed informed consent.

Pre-registration

The pre-registration for time point one (participants in rehabilitation unit) involved correlating OxMET scores with admission and discharge functional outcome measures (pre-registration https://osf.io/z37h8). Sample size was determined a priori by the smallest effect size of interest which was a correlation of $r = .30$, a power of 80%, and an alpha of .05, which lead to a need for a minimum of 66 individuals. The smallest effect size of interest was determined due to being a correlation benchmark for convergence (Rotenberg et al., 2020; Webb et al., 2022). We aimed to recruit as many participants as possible at time point one (recruitment in hospital) until we *reached* 66 participants at follow up, to account for expected attrition (previous inclusive recruitment of moderate stroke has shown up to 50% attrition, e.g., see Demeyere et al., 2021).

We deviated from our pre-registration variables but not hypotheses: we only correlated outcome measures taken closest to administration of the OxMET rather than specifically at discharge or admission to ensure complete data. We added disability outcome measures to examine different aspects of function that were clinically relevant. We added in an additional analysis with Montreal Cognitive Assessment (MoCA) data – where available – to investigate the specificity of the relationship between activities of daily living with the OxMET in comparison to a global cognitive screen.

Measures

Multiple functional outcome measures from basic independence to more holistic functioning including well-being were included based on clinical recommendation from the stroke rehabilitation unit.

The Oxford digital multiple errands test (OxMET)

The OxMET (Webb et al., 2022) is a computer-tablet based task requiring participants to buy specific items and answer questions on a list by going through shops on a screen, and to do so following clear constraints on time, money, and shop revisits. The OxMET is explained in detail in the original manuscript (Webb et al., 2022), which includes images for reference. The primary outcome metric is accuracy of task completion which ranges from −10 to 10 depending on rules followed and tasks completed per shop (e.g., not to enter a distractor shop, to enter and correctly purchase the items in item shops, and to enter and correctly answer the questions in question related shops). Additional metrics on error types and timing are available too (see Webb et al., 2022). To alleviate communication difficulties, with the help of the rehabilitation unit's Speech and Language Therapy team, visually presented standardized instructions, which are aphasia friendly, were created and used. These instructions are now used in the freely available app version. The OxMET is available to professionals via a free license from the PlayStore and App Stores.

The OxMET is a research tool and is not regulated as a medical device, as it does not provide a clinical diagnosis. The OxMET does not store personally identifiable data, avoiding any risks of data breaches and as such it fits with technology information safety guidance from the APA (APA taskforce on psychological assessment and evaluation guidelines, 2020). Moreover, documentation is provided alongside the OxMET for end users regarding the use and interpretation of the OxMET and guidance for use by non-experts (Bauer et al., 2012).

Functional outcome measures collected at time point 1 (baseline)

We used the Barthel Index as a commonly used and validated tool to classify basic independence (Mahoney & Barthel, 1965; Quinn et al., 2011). We used the stroke specific Therapy Outcomes Measure (Enderby et al., 2013)(TOMS) as a more holistic measure of functioning, on recommendations of Occupational Therapists at the unit. The TOMS examines four domains each rated by a physio – or occupational-therapist: Impairment (i.e., physical, cognitive, sensory impairment etc.), Activity (i.e., how able or independent one is at purposeful action), Participation (i.e., ability to involve oneself in their own life, be that socially, emotionally, educationally etc), and Well-being / Distress (i.e., scales from uncontrollably distressed/emotional to appropriate levels of well-adjusted emotionally) (Enderby et al., 2013, pp. 207–209). Each domain is scored on a 11-point scale (0 being the worst and 5 being the best, with half-points used to indicate slight deviation from the integer categories) (Enderby et al., 2013, p. 23). We created a summed total score of the TOMs as an outcome measure to compare to OxMET performance and reduce multiple comparisons.

The modified Rankin Scale (mRS) score at discharge was collected from clinical notes. The mRS is the most widely used stroke outcome measure in acute stroke clinical trials (Broderick et al., 2017), and it measures global disability on a scale of 0 = no symptoms to 6 = death.

Finally, The National Institute of Health Stroke Scale (NIHSS) rates stroke severity on a 15-point scale covering consciousness, orientation, vision, physical ability, language and speech, sensation, visual spatial attention, and changes since previous examination or baselines (Brott et al., 1989). The NIHSS taken closest to OxMET assessment was used as a covariate to partial the effect of stroke severity on associations.

Functional outcome measures collected at time point 2 (follow up)

Participants who completed follow-up, completed cognitive screening with the Oxford Cognitive Screen (Demeyere et al., 2015) and OCS-Plus (Demeyere et al., 2021) and the Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2005), and were administered a series of questionnaires as part of the OCS-RECOVERY protocol. In the current study, we only examined the Nottingham Extended Activities of Daily Living (NEADL) (Nouri & Lincoln, 1987) 66pt scale (Harwood & Ebrahim, 2002), the MoCA, and the mRS for disability and dependence level as our outcome measures.

Participants

A total of 117 participants were recruited to this study, with 66 stroke survivors completing the 6 months follow up. Attrition until follow up occurred due to: loss of contact (n = 19), death before discharge (n = 2), death before follow up (n = 10), no interest in completing the follow up (n = 7), too unwell to complete follow up within 2 month time frame (n = 6), and that we reached 66 participants before their follow up was due (n = 7). Participants were recruited and followed up between November 2020 and December 2022. We did not record eligibility screening data.

Inclusion criteria were: confirmed stroke; ability to give informed consent for participation in the study; age 18 or above; well enough to concentrate for 20 mins (as judged by the multi-disciplinary team in the hospital); and sufficient English to comprehend the initial orienting questions in the Oxford Cognitive Screen (OCS) (Demeyere et al., 2015). To avoid researcher bias, researchers approached all eligible participants the MDT deemed fit enough to participant on the day of recruitment. Exclusion criteria were visual or hearing impairments which prevented the patient from adequately seeing the task screen or hearing the instructions.

We present the summary demographics of the sample in Table 1. Missing NIHSS, pre-morbid mRS, and mRS at follow up values were imputed using the mean for each variable respectively, proportion of missingness is presented in Table 1.

Table 1. Summary of sample characteristics for all participants and those followed up 6 months post-original consent.

	All participants		Followed up	
Characteristic	N (missing %)	Values	N (missing %)	Values
Age (*M*(*SD*))	117 (0%)	74.44 (12.88)	66 (0%)	73.94 (12.68)
Education (*M* (*SD*))	117 (0%)	13.13 (3.25)	66 (0%)	13.52 (3.47)
Handedness	117 (0%)	R: 88.03%; L: 11.97%	66 (0%)	R: 81.82%; L: 18.18%
Sex	117 (0%)	M: 56.41%; F: 43.59%	66 (0%)	M: 60.61%; F: 39.39%
Ethnicity	117 (0%)	White-British: 90.60%; White-Other: 5.98%; Black-African: 1.71%; Black-Caribbean: 0.85%; Other-Asian: 0.85%	66 (0%)	White-British: 92.42%; White-Other 4.55%; Black African: 1.52%; Black-Caribbean: 1.52%
NIHSS (*Median* (IQR))	117 (0%)	8.32 (5-11)	66 (0%)	8 (4-11)
Time Since Stroke (Days) (*M*(*SD*))	117 (0%)	26.18 (25.16)	66 (0%)	201.76 (26.82)
Stroke type	117 (0%)	ischaemic: 80.35%; intracerebral haemorrhage: 14.53%; ischaemia with haemorrhagic transformation: 4.27%; subarachnoid haemorrhage: 0.85%	66 (0%)	ischaemic: 74.24%; intracerebral haemorrhage: 19.70%; ischaemia with haemorrhagic transformation: 4.55%; subarachnoid haemorrhage: 1.51%
Stroke side	117 (0%)	R: 52.99% L: 29.91%; B: 17.09%	66 (0%)	R: 53.03% L: 30.3%; B: 16.67%
Language Disturbance	116 (1%)	None: 60.34%; Dysarthria: 17.24%; Expressive Aphasia: 13.79%; Mixed Dysphasia (Expressive + Receptive): 6.9%; undetermined: 1.72%;	65 (2%)	None: 66.15%; Dysarthria: 15.38%; Expressive Aphasia: 10.77%; Mixed Dysphasia (Expressive + Receptive): 6.15%; undetermined: 1.54%;

In Figure 1, we present the distributions for each of the functional outcome measures for context of the sample. Figure 1 shows that participants were moderately disabled on the mRS, had moderate strokes on the NIHSS, were mild ranging on the Barthel Index and there was a wide range of holistic functioning scores on the TOMS. This describes the range in participants abilities, demonstrating a realistic representation of an in-patient rehabilitation setting.

Procedure

Participants were approached by a member of the team (SSW) at bedside in the rehabilitation unit after discussion from the MDT about inclusion/exclusion criteria. Following informed consent, participants were screened with the OCS and if there was enough time before other therapy or discharge, they were administered the OxMET. The Barthel, TOMS, NIHSS, mRS, and sometimes the OCS, were administered by the clinical team separately. Clinical and some demographic information were collected from medical notes. Participants were followed up 6 months later in person at home. Questionnaires were sent in advance via post.

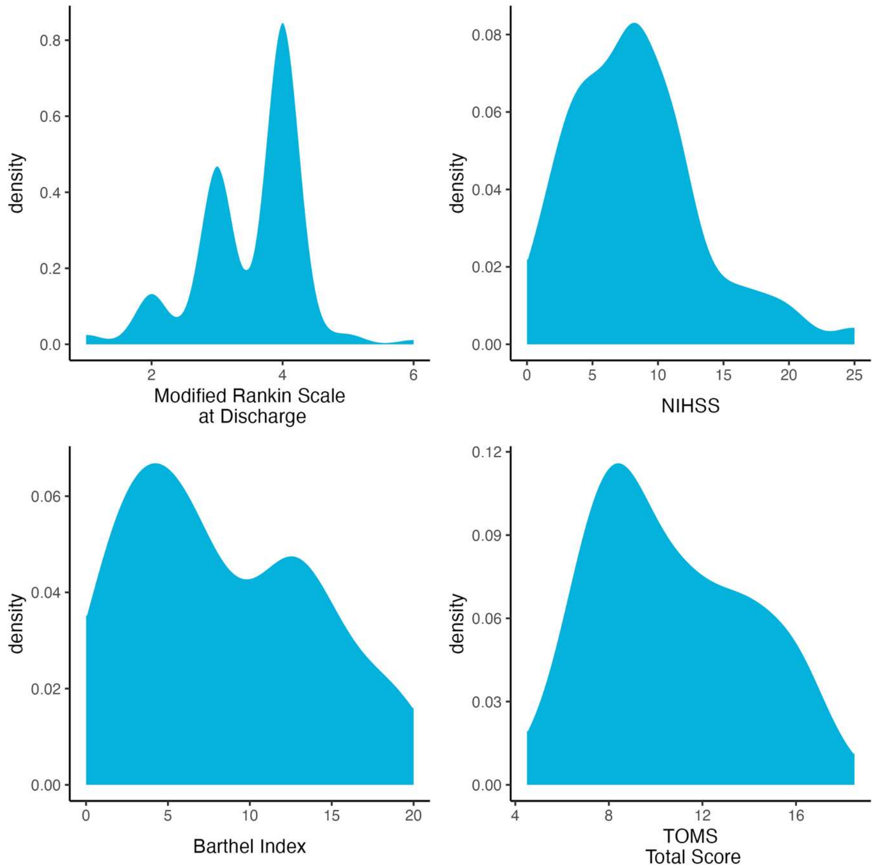

Figure 1. Illustrates the density distributions of functional outcome measures from the stroke sample of 117 individuals (<3 months post-stroke). Figure available by CC BY 4.0 attribution license https://osf.io/fqc6g.

Data analysis

As pre-registered, we examined correlations between functional measures taken in the rehab setting to OxMET metrics through Spearman's Rho correlation to establish any relationships. Next, we looked at the predictive abilities of OxMET performance in the rehab setting to activities of daily living 6-months following study participation. A regression analysis of the OxMET metrics and outcomes of the NEADL total score was conducted, with covariates of change in mRS score, NIHSS, age, and education. We were powered to run the regression analysis on a minimum of 66 participants. We conducted these regression models to investigate the potential predictive ability of OxMET accuracy. The OxMET accuracy score was chosen as it is the main outcome metric and intended to simplify impairment classifications and interpretation. As we only collected other cognitive data at 6-months, we compared the prediction of the NEADL score by OxMET score, to the prediction of NEADL score by

6-month MoCA score, using the same covariates. We then included the 6-month MoCA in the baseline OxMET prediction model of 6-month NEADL to assess the independent contributions of both measures. For correlations and regressions we used a convergence benchmark of $r = .30$, with acceptability of correlations above .19 as per other very commonly used neuropsychological batteries (see Demeyere et al., 2021 which discussed convergence acceptibility).

Data wrangling and statistical software

Most data wrangling was completed in MATLAB (The MathWorks Inc, 2021). All statistical analysis and some data sorting were computed in R studio (R Core Team, 2021) version 4.2.1. We used the following packages for the production of the RMarkdown manuscript and analysis: *bookdown* version 0.27 (Xie, 2021); *tidyverse* version 1.3.2 (Wickham et al., 2019); *readxl* version 1.4.0 (Wickham & Bryan, 2019); *cowplot* version 1.1.1 (Wilke, 2020); *ggplot2* version 3.3.6 (Wickham, 2016); *kableExtra* version 1.3.4 (Zhu, 2021); *httr* version 1.4.3 (Wickham, 2022); *ggsignif* version 0.6.3 (Ahlmann-Eltze & Patil, 2021); *Hmisc* version 4.7-0 (Harrell Jr, 2021); *pROC* version 1.18.0 (Robin et al., 2011); *data.table* version 1.14.2 (Dowle & Srinivasan, 2021); and *gridExtra* version 2.3 (Auguie, 2017).

Results

Associations with baseline outcome measures

Accuracy, the main outcome metric from the OxMET, was significantly related to the TOMS total score ($r(97) = .27$, $p = .007$) and mRS ($r(114) = -.30$, $p = .001$) below a multiple comparison corrected alpha ($p < .05/3$ comparisons), and the Barthel Index ($r(77) = .23$, $p = .04$). The correlation coefficient for the mRS was at our benchmark for convergent correlation, and the Barthel Index and TOMS total score was within our acceptable correlation criteria. Correlations between secondary OxMET metrics and the functional measures are presented in Table S1 in supplementary materials.

Predicting longer-term functional outcomes

Baseline OxMET accuracy was a significant predictor of NEADL total score ($R^2 adjusted$ = .29, $F(5,57) = 5.98$, $p<.001$, $\beta_1 = 2.1$, $p<.001$), when accounting for variance explained by change in disability and dependence (change in mRS score), age, education, and stroke severity at baseline (NIHSS score), and with a correction for multiple comparisons at the level of .05/4 ($p < .013$) . We found that OxMET accuracy demonstrated a clear positive predictive association with the NEADL. The simple associations between OxMET accuracy and NEADL scores are presented in Figure 2.

Finally, a linear regression with NEADL score predicted by MoCA total score taken at 6-months was completed using the same covariates and we found no significant association with MoCA, only with stroke severity ($R^2 adjusted = .21$, $F(5,43) = 3.535$, $p = .009$, $\beta_{1MoCA} = 0.7$, $p = .06$). When we included the 6-month MoCA in the OxMET prediction model, only the OxMET accuracy metric was a significant predictor of NEADL score at 6-months ($R^2 adjusted = .34$, $F(6,42) = 5.03$, $p < .001$, $\beta_{OxMET} = 1.88$, $p = .004$, $\beta_{MoCA} = 0.44$, $p = .251$). MoCA was only available for 50 out of 66 participants, and the differing sample sizes for the models will have affected the power to detect significant effects. However, we also note that the strength of the association increased for the OxMET with the second model including MoCA.

Discussion

We aimed to provide evidence for the predictive validity of the Oxford Digital Multiple Errands Test (OxMET) using a subacute (<2 months post-stroke) stroke survivor sample seen at two time points (baseline in a stroke rehabilitation unit and then 6 months later at home). The main analysis of the current

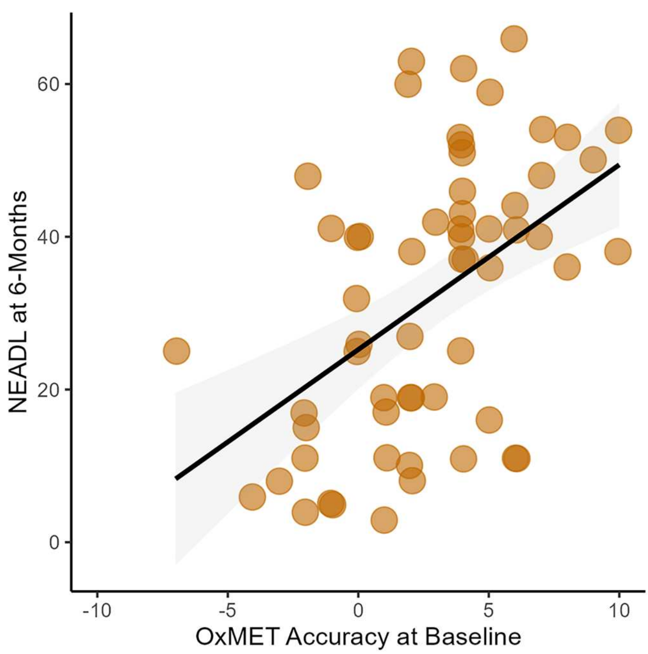

Figure 2. Visually jittered scatter plots between OxMET Accuracy scores and functional outcome measure on the Nottingham Extended Activities of Daily Living Scale (NEADL) taken at 6-month follow up. Figure available by CC BY 4.0 attribution license https://osf.io/jsbgy.

study concerned associations between subacute OxMET and functional abilities at baseline and at 6-month follow up. We previously demonstrated the OxMET has good psychometric validity (Webb et al., 2022) and converges with an in-person MET (Webb et al., 2022). Here we further evidence the case for usefulness of the OxMET by demonstrating predictive validity.

The OxMET significantly associated with the Therapy Outcomes Measure total score (TOMS) (Enderby et al., 2013) and modified Rankin Scale (mRS) (Broderick et al., 2017), but not the Barthel Index (Mahoney & Barthel, 1965), when correcting for multiple comparisons. However, the Pearson's r did not breach the threshold for convergence ($r >= .30$) for the TOMS, and as such we can only cautiously interpret this association. For 6-month follow up, we found a moderately significant association between OxMET accuracy and Nottingham Extended Activities of Daily Living (NEADL) scale (Nouri & Lincoln, 1987), over and above variance explained by stroke severity, demographics, and disability change. These findings suggest the OxMET, as an executive functioning screening tool, can be used as a valid assessment with a good degree of predictive validity to functional outcomes in activities of daily living. Nevertheless, inferences about real life functioning should still not be made on an individual prediction level.

Furthermore, we found a significant predictive relationship between OxMET and 6-month NEADL score, but we did not find one for concurrent MoCA score and NEADL. It remains possible that there could have been a predictive relationship from baseline MoCA to future functioning, but we did not collect baseline MoCA scores here. Indeed others have found such a relationship (Abzhandadze et al., 2019; Zietemann et al., 2018) and we previously also demonstrated a relationship between MoCA and OxMET (Webb et al., 2022). However, here we found a specific relationship from baseline OxMET, which could not be explained by concurrent MoCA scores.

There is limited research regarding the association between variants of the Multiple Errands Test (MET) and basic functional outcome measures relevant for rehabilitation settings. The Barthel index score is commonly used as an exclusion/inclusion criterion rather than as an outcome measure due the physical nature of the MET (Burns et al., 2019; Morrison et al., 2013; Raspelli et al., 2012). As such it is difficult to compare our OxMET and other METs in their relation to basic functional outcome measures. However, other more complex functional measures have been investigated in comparison to METs. For example, Dawson et al. (2009) compared their Baycrest MET version to assessments of motor and cognitive impairments and complex activities of daily living in stroke survivors who were younger, more educated, and more chronic (8.6years post-stroke) than our sample (Dawson et al., 2009). They found significant moderate to large correlations above .30 between their error scores, but not accuracy. Maeir et al. (2011) investigated the associations of errors in the hospital version of the MET (Knight et al., 2002) with participation in the community 3-months post-stroke (Maeir et al., 2011). The study found correlations above .30 for most MET metrics except task failures, and they did not use an accuracy metric. When we re-ran regressions including the error OxMET metrics, we found that accuracy was still the only metric related to activities of daily living, over and above change in mRS, NIHSS, age, and education. This strengthens our justification for using an overall outcome metric, to make clinical interpretation easier. Differences between results may be due to our sample, which was more severely affected by their stroke, older, and with greater levels of disability.

Study limitations

There are limited tools which can measure functional ability in a standardized and psychometrically validated way in in-patient settings. We were reliant therefore on functional assessments which revolve around basic ADLs, such as the Barthel, or assessments which are confounded by physical impairments. This is a limitation in that these tasks cannot reflect real life behaviours which are complex. However, the Barthel and TOMS are commonly used in clinical practice and provide an easily interpretable scale to compare to the relatively new OxMET task.

Separately, our study deviated from the standardized administration of the OxMET in that we used supplemental instructions developed with the speech and language team to aid comprehension of instructions for participants with lower communication abilities. The use of our images paired with broken down and spaced-out text has now been implemented in the latest version of the OxMET app available on the appstores. The small differences in administration could have conceivable led to higher performance, though arguably also allow for the interpretation of a more specific executive deficit, rather than a potential confound from a problem in complex instruction comprehension. There was no evidence of an overall better performance though, with the current stroke cohort as a whole scoring well below the healthy ageing data (Webb et al., 2022). Updated normative data will continue to be added to the app (some new normative data is also presented in Webb et al., 2022) and we continue to provide open data for readers interested in direct comparison.

We recruited only from a specialist stroke in-patient rehabilitation unit and not from acute or hyper acute, or out-patient settings, restricting generalizability. We feel that this is countered by using a moderate stroke sample, where most research uses mild stroke cohorts.

Future directions

The OxMET has been psychometrically validated as an executive functioning screening tool, and now has evidence to support its predictive validity. Future acceptability and feasibility of the OxMET in clinical practice, administered by allied health professionals, will need to be formally assessed.

Conclusion

The OxMET total score associated with basic functional measures and predicted activities of daily living. Its associations with clinically relevant and activity relevant measures provide key evidence to support its use in real-world clinical settings.

Data availability

The data that support the findings of this study are openly available in Open Science Framework at http://doi.org/10.17605/OSF.IO/WEP3V.

Disclosure statement

Nele Demeyere is a developer of the "OxMET" but does not receive any remuneration from its use.

Funding

This work was supported by Health Services and Delivery Research Programme [Grant Number Advanced Fellowship NIHR302224]; Stroke Association [Grant Number SA PGF 21100015].

ORCID

Sam S. Webb http://orcid.org/0000-0002-0029-4665
Nele Demeyere http://orcid.org/0000-0003-0416-5147

References

Abzhandadze, T., Rafsten, L., Lundgren Nilsson, Å, Palstam, A., & Sunnerhagen, K. S. (2019). Very early MoCA Can predict functional dependence at 3 months after stroke: A longitudinal, cohort study. *Frontiers in Neurology*, *10*. https://www.frontiersin.org/articles/10.3389fneur.2019.01051

Ahlmann-Eltze, C., & Patil, I. (2021). *ggsignif: Significance Brackets for 'ggplot2'* [Manual]. https://CRAN.R-project.org/package = ggsignif

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

American Educational Research Association. (2014). American Psychological Association, & National Council on Measurement in Education. In *Standards for educational and psychological testing* (pp. 11–47). American Educational Research Association.

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*. https://doi.org/10.3389/fpsyg.2019.02575

APA taskforce on psychological assessment and evaluation guidelines. (2020). *APA Guidelines for Psychological Assessment and Evaluation*. https://doi.org/10.1037/e510142020-001

Auguie, B. (2017). *gridExtra: Miscellaneous functions for 'Grid' graphics* [Manual]. https://CRAN.R-project.org/package = gridExtra

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

Broderick, J. P., Adeoye, O., & Elm, J. (2017). Evolution of the modified rankin scale and Its Use in future stroke trials. *Stroke*, *48*(7), 2007–2012. https://doi.org/10.1161/STROKEAHA.117.017866

Brott, T., Adams, H. P., Olinger, C. P., Marler, J. R., Barsan, W. G., Biller, J., & Hertzberg. (1989). Measurements of acute cerebral infarction: a clinical examination scale. *Stroke*, *20*(7), 864–870. https://doi.org/10.1161/01.STR.20.7.864

Burns, S. P., Dawson, D. R., Perea, J. D., Vas, A., Pickens, N. D., & Neville, M. (2019). Development, reliability, and validity of the multiple errands test home version (MET–home) in adults With stroke. *The American Journal of Occupational Therapy*, *73*(3). https://doi.org/10.5014/ajot.2019.027755

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 Parkinsonâ€™s disease. *Frontiers in Behavioral Neuroscience*, *8*, 405. https://doi.org/10.3389/fnbeh.2014.00405

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. J., Wright, H., Finke, K., & Duta, M. (2021a). Introducing the tablet-based Oxford cognitive screen-plus (OCS-plus) as an assessment tool for subtle cognitive impairments. *Scientific Reports*, *11*), https://doi.org/10.1038/s41598-021-87287-8

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. https://doi.org/10.1037/pas0000082

Demeyere, N., Williams, O. A., Milosevich, E., Chiu, E. G., Drozdowska, B. A., Dillon, A., Dawes, H., Thomas, S., Kuppuswamy, A., & Pendlebury, S. T. (2021b). Long-term psychological consequences of stroke (OX-CHRONIC): A longitudinal study of cognition in relation to mood and fatigue after stroke: Protocol. *European Stroke Journal*, *6*(4), 428–437. https://doi.org/10.1177/23969873211046120

Dowle, M., & Srinivasan, A. (2021). *data.table: Extension of `data.frame`* [Manual]. https://CRAN.R-project.org/package = data.table

Enderby, P., John, A., & Petheram, B. (2013). *Therapy outcome measures for rehabilitation professionals: Speech and language therapy, physiotherapy, occupational therapy*. John Wiley.

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

Harrell Jr, F. E. (2021). *Hmisc: Harrell Miscellaneous* (R package version 4.5.0) [Computer software]. https://CRAN.R-project.org/package = Hmisc

Harwood, R. H., & Ebrahim, S. (2002). The validity, reliability and responsiveness of the Nottingham extended activities of daily living scale in patients undergoing total hip replacement. *Disability and Rehabilitation*, *24*(7), 371–377. https://doi.org/10.1080/09638280110101541

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

Katz, N., Tadmor, I., Felzen, B., & Hartman-Maeir, A. (2007). The behavioural assessment of the dysexecutive syndrome (BADS) in schizophrenia and its relation to functional outcomes. *Neuropsychological Rehabilitation*, *17*(2), 192–205. https://doi.org/10.1080/09602010600685053

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

Lipskaya-Velikovsky, L., Zeilig, G., Weingarden, H., Rozental-Iluz, C., & Rand, D. (2018). Executive functioning and daily living of individuals with chronic stroke: Measurement and implications. *International Journal of Rehabilitation Research*, *41*(2), 122–127. https://doi.org/10.1097/MRR.0000000000000272

Maeir, A., Krauss, S., & Katz, N. (2011). Ecological validity of the multiple errands test (MET) on discharge from neurorehabilitation hospital. *OTJR: Occupation, Participation and Health*, *31*(1_suppl), S38–S46. https://doi.org/10.3928/15394492-20101108-07

Mahoney, F., & Barthel, D. W. (1965). Functional evaluation ; the barthel index. A simple index of the independence useful in scoring improvement in the rehabilitation of the chronically ill. *Maryland State Medical Journal*, *14*, 61–65.

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

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

Morrison, M. T., Giles, G. M., Ryan, J. D., Baum, C. M., Dromerick, A. W., Polatajko, H. J., & Edwards, D. F. (2013). Multiple errands test–revised (MET–R): A performance-based measure of executive function in people with mild cerebrovascular accident. *The American Journal of Occupational Therapy*, *67*(4), 460–468. https://doi.org/10.5014/ajot.2013.007880

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Nouri, F. M., & Lincoln, N. B. (1987). An extended activities of daily living scale for stroke patients. *Clinical Rehabilitation*, *1*(4), 301–305. https://doi.org/10.1177/026921558700100409

Oksala, N. K. J., Jokinen, H., Melkas, S., Oksala, A., Pohjasvaara, T., Hietanen, M., Vataja, R., Kaste, M., Karhunen, P. J., & Erkinjuntti, T. (2009). Cognitive impairment predicts poststroke death in long-term follow-up. *Journal of Neurology, Neurosurgery & Psychiatry*, *80*(11), 1230–1235. https://doi.org/10.1136/jnnp.2009.174573

Pohjasvaara, T., Leskelä, M., Vataja, R., Kalska, H., Ylikoski, R., Hietanen, M., Leppävuori, 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. J., Langhorne, P., & Stott, D. J. (2011). Barthel index for stroke trials. *Stroke*, *42*(4), 1146–1151. https://doi.org/10.1161/STROKEAHA.110.598540

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.

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. https://doi.org/10.1080/09602010802469074

Raspelli, S., Pallavicini, F., Carelli, L., Morganti, F., Pedroli, E., Cipresso, P., Poletti, B., Corra, B., Sangalli, D., & Silani, V. (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. (2021). *R: A language and environment for statistical computing*. [Computer Software]. R Foundation for Statistical Computing,. https://www.R-project.org/

Reitan, R. M., & Wolfson, D. (1995). Category test and trail making test as measures of frontal lobe functions. *The Clinical Neuropsychologist*, *9*(1), 50–56. https://doi.org/10.1080/13854049508402057

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. https://doi.org/10.1186/1471-2105-12-77

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*, *101*(9), 1628–1642. https://doi.org/10.1016/j.apmr.2020.01.019

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.2160588

The MathWorks Inc. (2021a). *MATLAB and statistics toolbox* [Computer software]. The MathWorks, Inc.

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *The Lancet*, *370* (9596), 1453–1457. https://doi.org/10.1016/S0140-6736(07)61602-X

Webb, S. S., Anders, J., Chiu, E. G., Payne, F., Basting, R., Duta, M. D., & Demeyere, N. (2022). The Oxford digital multiple errands test (OxMET): Validation of a simplified computer tablet based multiple errands test. *Neuropsychological Rehabilitation*, https://doi.org/10.1080/09602011.2020.1862679

Wickham, H. (2016). *Ggplot2* (2nd ed). Springer International.

Wickham, H. (2022). Httr: Tools for Working with URLs and HTTP (R package version 1.4.2) [Computer software]. https://CRAN.R-project.org/package = httr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Bryan, J. (2019). *readxl: Read excel files* (R package version 1.3.1) [Computer software]. https://CRAN.R-project.org/package = readxl

Wilke, C. O. (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'* (R package version 1.1.1) [Computer software]. https://CRAN.R-project.org/package = cowplot

Wilson, B. A. (1993). Ecological validity of neuropsychological assessment: Do neuropsychological indexes predict performance in everyday activities? *Applied and Preventive Psychology*, *2*(4), 209–215. https://doi.org/10.1016/S0962-1849(05)80091-5

Wilson, B. A., Evans, J. J., Alderman, N., Burgess, P. W., & Emslie, H. (1997). Behavioural assessment of the dysexecutive syndrome. *Methodology of Frontal and Executive Function*, *239*, 250.

Xie, Y. (2021). *bookdown: Authoring books and technical documents with r markdown* [Manual]. https://github.com/rstudio/bookdown

Zhu, H. (2021). *kableExtra: Construct complex table with 'kable' and pipe syntax* [Manual]. https://CRAN.R-project.org/package = kableExtra

Zietemann, V., Georgakis, M. K., Dondaine, T., Müller, C., Mendyk, A.-M., Kopczak, A., Hénon, H., Bombois, S., Wollenweber, F. A., Bordet, R., & Dichgans, M. (2018). Early MoCA predicts long-term cognitive and functional outcome and mortality after stroke. *Neurology*, *91* (20), e1838–e1850. https://doi.org/10.1212/WNL.0000000000006506