NeuroRehabilitation 35 (2014) 553–562
DOI:10.3233/NRE-141151
IOS Press

Using video ratings to assess multitasking performance in a naturalistic paradigm

S.M. Hynes[a,c,*], J. Fish[b,c] and T. Manly[c]
[a]*Dementia Research Centre, North East London NHS Foundation Trust, London, UK*
[b]*King's College London, Institute of Psychiatry, Psychology & Neuroscience, Department of Psychology, North East London NHS Foundation Trust, London, UK*
[c]*MRC Cognition and Brain Sciences Unit, Cambridge, UK*

Abstract.
BACKGROUND: Multitasking measures, in which a series of tasks must be completed within a naturalistic setting not fully under the experimenter's control, have been shown to be more sensitive than traditional measures in detecting organisational problems in people with difficulties in executive functioning. There are a number of drawbacks to such tasks however. They can take considerable time to administer and are demanding in terms of examiners noting and recording all relevant aspects of performance. This potentially leaves them more open to subtle bias. One method that could offset these limitations is to video record performance.
OBJECTIVES: The practicality and outcome of using video ratings to accurately score performance off-line is investigated here.
METHODS: Nineteen participants completed a Multiple Errands Task (MET) while wearing a body-worn camera. Their performance was scored "live" and by an independent rater who had only access to video footage of the task.
RESULTS: Significant relationships were seen on all variables of the MET between the live and video ratings. The inter-rater reliability of the measure appears strong.
CONCLUSION: We provide initial support for the use of a video rater when assessing performance on an MET.

Keywords: Rehabilitation, memory; neurorehabilitation, assessment, therapy, executive function

1. Introduction

Psychologists often try to develop highly controlled tasks to isolate a particular cognitive capacity and to minimize the effect of different prior experience by using novel, abstract materials – the Wisconsin Card Sorting Test is a good example. Here people are asked to sort a pack of cards showing shapes of different colours in different groupings according to logical rules (by colour, shape or number of items; Heaton, 1981). Every so often the participant is asked to switch the rule. The task is conducted under quiet one-to-one conditions with the examiner indicating when to start and stop the task. From this, inferences are drawn about people's ability to hold on to and switch mental set with presumed predictive validity for their abilities in everyday settings that require these skills. It has been noted by a number of authors (e.g. Shallice & Burgess, 1991) that such traditional desk-top measures may in fact be rather *insensitive* to executive problems that are manifest in normal situations. They argue that when you reduce a task to assess a specific capacity you throw out many features with which such people may struggle. A key feature of everyday situations is that we generally have *multiple* goals which, due to our capacity to only complete one at a time, are in competition with each other (planning tomorrow's packed lunch is in competition with doing the washing up which is in competition with watching the TV). This delay in being able to act on a

*Address for correspondence: Dr. Sinéad Hynes, Dementia Research Centre, Research & Development Department, North East London NHS Foundation Trust, 1st floor, Maggie Lilley Suite, Goodmayes Hospital, Barley Lane, Ilford, Essex IG3 8XJ, UK. Tel.: +44 0300 555 1200/Ext: 4491; E-mail: sinead.hynes@nelft.nhs.uk.

goal can lead to it being forgotten. Similarly, everyday life contains many habitual triggers for actions that we may not even consciously intend to complete (I may go into a shop to buy eggs but return with some washing powder that was on special offer and forget all about the eggs). Particularly where there is insufficient time to complete all of one's goals it is necessary to prioritise and review 'on the hoof' in a way that takes into account the opportunities and barriers that one encounters. Accordingly researchers have attempted to develop tasks that build in rather than exclude these features.

1.1. The development of the MET

In 1991 Shallice and Burgess developed a measure, the Multiple Errands Test (MET), with this aim in mind. Participants were asked to complete a series of goals within a given section of a London shopping street. These included buying specific items and finding out information. Some general knowledge and inference was required (e.g. how you could find which part of the UK had been hottest on the previous day). Participants also had to comply by task-specific rules (such as not re-entering a shop) as well as socially normative and legal rules (e.g. not insulting shop staff or stealing items). Planning was required to develop a strategy likely to complete the tasks within a given time (e.g. ordering the tasks to minimize the distance that needed to be walked) and this plan had to be held in mind and, if necessary, updated over the period of performance. The participant's behaviour was carefully recorded by an examiner who followed them at a distance, who was also ready to intervene if rule breaking became problematic! Shallice and Burgess (1991) showed that, in three patients with frontal lobe deficits who performed well on tests of IQ, perception, language and cognition, the MET elicited the types of errors that were apparent in their everyday lives. They committed a large number of rule breaks and developed inefficient strategies. They had problems with the task because it focused on areas of organisation and managing multiple sub-goal tasks that had been shown to be difficult for them. The task also requires motivation and memory, and healthy control participants do not often score at ceiling. From this paper it is clear, despite the small sample involved, that a multiple sub-goal-type task might be an appropriate measure to use with people who score within the normal level on other psychometric tests.

Although it does not make reference to the term "MET", Boyd and Sautter (1993) developed a similar unstructured task that could be used with a brain-injured population. Participants' ($n = 31$) route-finding ability was assessed in a hospital setting where the raters looked at task formulation, strategy application, their dependence on cueing and the detection and correction of mistakes. In this task, as with the MET, participants were free to complete the task in any number of ways. Unlike the point-by-point recording of the MET, however, a Likert scale was used to rate overall performance. Although the authors cite good inter-rater reliability, this single score raises interesting questions about the precise criteria used by different raters. Burgess et al. (2000) tackled this issue in a revised MET that attempted to break down performance into theoretically separable categories. For example, asking people about their plan before and after performance allows planning to be separated from memory for, and tendency to follow, the plan.

Since the initial development of the MET a number of versions have been reported for use with clients with more severe difficulties. A hospital-based version of the MET (MET-HV) was designed by Knight, Alderman and Burgess (2002) for clients with behavioural problems. They assessed 20 healthy controls and 20 patients on this simplified version and found that patients showed most problems with subtle planning, prospective memory and when a task was "ill-structured". Patients broke more rules, made more mistakes, achieved fewer tasks and were more reliant on others to help them. They found that this version of the MET correctly classified 85% of patients. Knight et al. (2002) reported that in the MET-HV, particular failure to achieve tasks, combined with responses on the Dysexecutive Questionnaire (DEX; Burgess, 1996) gave an overall indication of the presence and severity of behavioural difficulties. They raise the important issue that failure on MET style measures can occur due to neglect of the plan but also frank amnesia for the tasks.

For patients with lower IQ but who can be safely assessed in a public setting Alderman, Burgess, Knight and Henman (2003) looked into a simplified version of the original MET (MET-SV). Patients had an IQ post-injury, as measured by the WAIS-R FSIQ, of 84.1 (SD12.7) and the majority (75%) were categorised as very severe traumatic brain injury, as determined by duration of post-traumatic amnesia and duration or depth of coma when first admitted to hospital. This version of the MET had simplified task demands, and more concrete rules, and had more time available for task completion. They reported on the basis of data from 46 controls and 50 patients that the key MET-SV

variables differentiating between the groups were the number of rule breaks and task failures. Patients made approximately three times more errors and significantly (19 times) more social rule breaks that healthy participants. As with previous studies it was noted that some of the patients who struggled with the MET had performed relatively well on traditional desktop measures of executive function.

1.2. Study aims

Clinicians are increasingly looking for assessments that are transparent to patients and that are representative of their needs and generalise to different settings. The MET lends to that possibility and has the potential to be accepted by both patients and clinicians. As described, there is good evidence that the complex, unstructured, multiple competing goals nature of MET, in that these mimic real life situations, can make the test better predictors of dysexecutive everyday errors than highly reduced/abstract desktop tests. A drawback to these tests is that this very complexity can make it difficult to clearly interpret errors and they are lengthy to administer. Finally, important limitations that this study seeks to address are that MET are hugely reliant on the attention of the administrator in noting what occurs and when and that, unlike many paper and pencil or computerised tests, the examiner's report provides the only available record. This is important because mistakes cannot be corrected. Here therefore the practicality and outcome of using first-person video recordings from a device worn by the participants was examined. Specifically, contemporary ratings from a "live" MET examiner were compared with those from an independent rater who scored during off-line viewing of video footage. The key questions were:

1. To what extent did the independent raters' scores accord with those of the live examiner?
2. Did the video recordings allow greater accuracy in some respect (e.g. events were noted that were missed by the live examiner, timings were more accurate etc.)?
3. Did the video reduce accuracy in some respects compared with the live rater (e.g. viewing perspective was non-optimal, technical glitches occurred etc.)?
4. Was the accuracy of the video recordings such that future examiners using this technology would be able to keep a watchful eye on participants' safety etc. but not concern themselves with live scoring?

2. Methodology

A variant of Shallice and Burgess' (1991) MET was developed to fit the layout and shops of Cambridge's Grafton Shopping Centre, and is described below. Participants drawn from the older healthy population were asked to wear a hidden video camera designed to continuously record sound and video from the participants 1st person perspective (i.e. the participants themselves were not seen but their voices, location, arm actions etc. should be visible). The choice of using a hidden video camera was to prevent attention being drawn to the participants, others behaving in unusual ways and/or having concerns about why filming was taking place. This required close consultation with the ethics committee and MRC Regulatory Statutory Support Unit about the legality of this type of filming. Crucially, shoppers and shop staff who appeared in our video recordings were incidental to the aims of the study. Ethical approval for this study was granted by Cambridge Psychology Research Ethics Committee (CPREC 2009.53). Permission to carry out the task in the Grafton Shopping Centre was granted in writing by its management. All participants gave informed consent both before and after taking part, were fully aware of the video camera and were reimbursed for their time.

Twenty-six participants from the older healthy population were recruited from the MRC Cognition and Brain Sciences Unit Volunteer Panel and carried out the video MET. Technical problems including poor video quality/angle, obscured camera angle, muffled audio and/or problems with the battery meant that seven of the original 26 videos could not be used. This left a sample size of 19 participants (13 male) with a mean age of 69.04 years (SD 5.22).

2.1. Pilot

Two versions of the MET were developed (see below) and piloted. If the task were to be used as an outcome measure in the future, having available a parallel version would help reduce the effects of practice and make the task more challenging and interesting for participants on re-assessment.

Convenience sampling was used for this small group. Four participants (one male; mean age 28.5 years, SD 10.47) took part in the pilot. Piloting took place in order to trial test procedures and to identify any potential difficulties with the two versions of the task. Some difficulties were identified, for example one shop that was to be used was closed down and there were difficulties

with positioning the camera. The timings of the two versions of the task were compared and seen to be of equal length and difficulty. Feedback from participants indicated that they found the task to be a reasonable challenge within the time and budget allowed. They also stated that the task would prove more difficult if it was to happen at a busier time as there would be longer queues and more distractions. As a result when possible the task was carried out before noon.

2.2. The MET

The MET was carried out in the Cambridge Grafton shopping centre, which was familiar to many participants. Before starting the participants donned the body-worn camera (CCD Button Camera). Initially a button camera was used attached to the body strap of a sports bag but camera angle and battery life were unreliable. A second camera (Swann Pen Cam™) was sourced and used. This widely available commercial product had wide-angle lens built into the lid of a pen, such that when the pen was clipped into a breast pocket it collected a stable, first person perspective view (see Fig. 1). Participants were asked to wear clothing with a breast pocket if possible. If not, solutions were improvised (such as clipping the pen to a cross-body bag strap).

Before being given the instructions, participants were asked two questions ("*How efficient would you say you were with tasks like shopping?*" "*How well do you know this shopping centre?*"). For the first question there was a 10-point response scale with end points labelled ("1" – "hopeless", "10" – "excellent"). The second had a four-point scale ("1" – "never visited", "2" – "visited once or twice", "3" – "visit occasionally", "4" – "visit regularly"). Participants were read and given the instructions on a piece of paper (with a clip board if participants wanted to use it) and given a pen, a plastic bag and a ten pound note. Participants were all asked to wear a wrist watch if they had one. If they did not then they were given a phone to carry with a clear clock display on the screen without having to press any buttons.

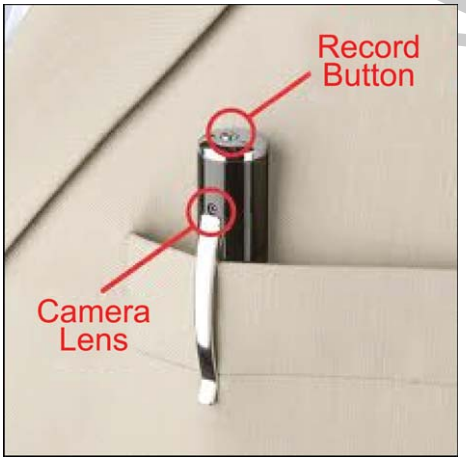

Fig. 1. The covert video recorder pen used in the MET.

Participants were first told the geographical limits of the shopping centre. The task instructions were as follows:

"In this exercise I want you to complete three tasks. The tasks are: to buy the five items listed on this sheet (*indicate and describe items on the sheet*); to obtain and write down five pieces of information (*indicate and describe items on sheet*); and to meet me here in 25minutes after I have said "… begin the exercise" and tell me the time. However, whilst completing this exercise you must obey the rules listed on your instruction sheet (*indicate and describe rules on sheet*).

You must carry out all these tasks but you may do so in any order. You should spend no more than £6; although I've given you £10 you should spend no more than six. You should stay within the limits of this shopping centre. You are free to go upstairs if you like but you must not go outside any of the outside doors. No shop should be entered other than to buy something, so if you go into a shop it should be with the intention of buying something. You should not go back into a shop you have already been in, so if you've been into a particular shop you should not go back into it again. You should only buy items from shops, not stalls. You should buy no more than two items from Poundland. Take as little time as possible to complete this exercise without rushing excessively.

Finally, approach me and tell me when you have completed the exercise.

Is that clear, have you any questions?" (*Clarify any questions the participant has*).

"Now tell me what you must do." (*Ensure participant is clear about what they must do*).

"*Begin the exercise*" (*Start timing at this point*).

Participants were given an instruction sheet to keep with them, which listed the rules and tasks. At the end of the test participants were asked to rate two more questions: "*How easy did you find the task*?" using a five-point scale with weighted end points ("1" – "very difficult", "5" – "very easy") and "*How well do you think you did with the shopping task?*" using a

Table 1
Examples of errors made across these categories

Task failure	Inefficiency	Rule break	Interpretation failure
Does not tell the time at the end	Does not use a time-efficient route during task	Speaks to the instructor during the task	Purchases sponges instead of dish cloths
Does not purchase an item or gather information necessary	Does not ask for help from shop assistant when it would be more efficient to do so	Leaves the outside limits of the shopping centre	Gets the phone-number of the wrong phone box

ten-point scale with weighted end points ("1" – "hopeless", "10" – "excellent"). Participants were randomly given either version A or version B. The instructions for both versions were the same. The only difference was in the information that was to be gathered and the items to be bought. A "live" rater followed each participant during the task and rated their performance from a set scoring sheet.

2.3. Scoring

Scoring of the task was in line with the categorisation of errors specified by Shallice and Burgess (1991, examples given in Table 1 above): inefficiencies, rule breaks, interpretation failures and task failures. An error was marked as an "*inefficiency*" when the participant could have used a different method to achieve the task more efficiently. "*Rule breaks*" apply to both the rules of the task and also social rules, for example shouting at a shop assistant. If a subtask was misunderstood it was deemed an "*interpretation failure*" and a "*task failure*" was when subtasks – buying items or collecting information – are not finished satisfactorily.

There was no maximum number of errors that participants could make. A participant received one point for every error they made – i.e. a lower score is a more efficient completion of the overall task. If participants performed well it was possible for them to make no errors.

The variables that were used in the MET task were "Tasks" which was made up of both the number of items bought and the pieces of information gathered (max. 10); "Mistakes" which was the number of errors made during the task including inefficiencies, rules breaks, interpretation failures and task failures; and "Time" which is the time it took to complete the MET in seconds. A record was also kept by the rater of the amount of money spent and an attempt at the time participants spent planning. This proved difficult, as will be discussed later.

2.4. Raters

The live ratings of the MET were scored as the task was taking place and directly on completion of the task. Two raters scored each video at a later time. The two video raters had no contact with each other and had no access to either the live ratings or the ratings of the other video rater. Video ratings were carried out at the rater's convenience in a quiet environment. The video raters were trained in scoring by the live examiner and used assessment sheets identical to those used in live scoring. Video raters were able to rewind, pause and replay any sections of the videos and to take their timings from the video clock.

3. Results

3.1. Timing

The participants were instructed to tell the examiner the time (from his or her watch) at which they had completed the test. The live examiner and video-raters recorded whether this had been achieved and the time that it occurred relative to the start of the test according to the stopwatch/video clock. Where participants did not remember to report the time the 'finish' point was set as 1500 (25 mins) and it was marked as a task failure. On three occasions, the video recorder was switched off by participants before this point. In this case the video-raters recorded the time the video ended.

As shown in Table 2, even when a margin of ± 10 seconds was used to take into account small differences in when timing started, rounding up etc., the raters only all agreed in 3/19 participants cases. This was also where the greatest variance was in performance between participants. As might be expected agreement was higher between the video raters but still only occurred in about half of cases. This level of agreement suggests that the scoring criteria were understood by the raters and

Table 2
Time to completion (seconds) scores from three raters (live, video rater 1 and video rater 2) for each of the participants MET performance

Participant	Live	Video 1	Video 2	All agree	Video raters agree	Overall discrep.	Video rater discrep.	L-V1 bias	L-V2 bias	V1-V2 bias
1	1500	1552	1500			52	52	−52	0	52
2	2280	2290	2340			60	50	−10	−60	−50
3	1560	1560	1500			60	60	0	60	60
4	1538	1600	1560			62	40	−62	−22	40
5	1570	1560	1500			70	60	10	70	60
6	1376	1380	1380	Y	Y	4	0	−4	−4	0
7	998	1020	1020		Y	22	0	−22	−22	0
8	1215	1380	1224			165	156	−165	−9	156
9	1218	1260	1260		Y	42	0	−42	−42	0
10	1288	1300	1260			40	40	−12	28	40
11	1656	1700	1630			70	70	−44	26	70
12	1500	1500	1500	Y	Y	0	0	0	0	0
13	1394	1440	1440		Y	46	0	−46	−46	0
14	1376	1370	1380	Y	Y	10	10	6	−4	−10
15	2295	2280	2220			75	60	15	75	60
16	1532	1500	1500		Y	32	0	32	32	0
17	1538	1530	1500			38	30	8	38	30
18	1451	1440	1440		Y	11	0	11	11	0
19	1479	1480	1490		Y	11	10	−1	−11	−10
Mean (SD)	1513.89	1533.79	1507.58	0.00	0.00	45.79	33.58	19.89	6.32	26.21
	(314.12)	(303.03)	(307.58)			(37.53)	(39.52)	(43.71)	(38.08)	(45.02)
Agree				15.70%	47.3%					

Agreement rates are taken as being within ±10 seconds of the other rater's score. Overall discrepancy represents the difference between the highest and lowest value reported by any rater. Video Discrepancy represents the difference between the highest and lowest values returned by the video raters. The bias scores take into account the direction of a discrepancy: L-V1 bias = Live rater – Video rater 1; L-V2 bias = Live rater – Video rater 2; V1–V2 bias = Video rater 1 – Video rater 2.

applied consistently where the events seen/filmed were unambiguous.

3.2. Tasks

Table 3 shows the range in performance on the MET variables as scored by the live rater, first and second video raters.

Looking at the scores presented in Table 4, there is perhaps a surprising lack of agreement in the number of tasks achieved by each participant. The raters' totals agreed in only 6/19 (32%) of cases between all three raters. The level was slightly higher between the two video raters (9/19, 47%). Examining these discrepancies, 10/13 (77%) was of 1 point or fewer but in two instances were 3 and 4 points. A likely case of non-random discrepancy would be if the achievement of some task was apparent to the live rater but obscured on the video. This would be consistent with the live rater's generally higher scores (see bias scores in Table 3). However, in one of these larger discrepancies, the live-rater (9 tasks complete) was in better agreement with the first video rater (8 tasks) than the second (5 tasks) suggesting either a period of inattention in the final rater or something rather ambiguous about this participant's performance. The reverse pattern was apparent in the second substantial discrepancy, with the live rater and video rater 2 being in greater agreement and video rater 1 noting markedly less task achievement. Unlike the purchasing of items it is often less clear to the videos raters if certain items of information have been collected. Some information, such as the number of the public phone box are more obvious on camera than other pieces of information such as the number of mobile phone shops in the centre or the number of shops beginning with a certain letter. It is less obvious through the video if participants are taking note of the number of shops or not – this is a lot clearer live. Some

Table 3
MET variables and participant performance($n = 19$)

Rater	Measure	Minimum	Maximum	Mean	Std. Deviation
Live	Tasks	7	10	9.08	0.75
	Time	998	2295	1513.89	314.115
	Mistakes	0	6	3.29	1.727
Video 1	Tasks	6	10	8.63	1.065
	Time	1020	2290	1533.79	303.032
	Mistakes	0	6	3.21	1.475
Video 2	Tasks	5	10	8.32	1.204
	Time	1020	2340	1507.58	307.434
	Mistakes	1	7	3.26	2.156

Table 4
Tasks achieved scores from three raters (live, video rater 1 and video rater 2) for each of participants MET performance

Participant	Live	Video 1	Video 2	All agree	Video raters agree	Overall discrep.	Video rater discrep.	L-V1 bias	L-V2 bias	V1–V2 bias
1	9	9	8			1	1	0	1	1
2	10	10	10	Y	Y	0	0	0	0	0
6	8	8	7			1	1	0	1	1
7	9	9	8			1	1	0	1	1
9	9	9	9	Y	Y	0	0	0	0	0
11	9	9	9	Y	Y	0	0	0	0	0
12	10	10	9			1	1	0	1	1
13	9	8	5			4	3	1	4	3
14	9	9	9	Y	Y	0	0	0	0	0
17	9	8	9			1	1	1	0	−1
20	10	10	8			2	2	0	2	2
24	7	7	7	Y	Y	0	0	0	0	0
25	10	9	10			1	1	1	0	−1
26	9	8	8		Y	1	0	1	1	0
27	9	9	8			1	1	0	1	1
28	9	8	8		Y	1	0	1	1	0
30	8.5	8	8		Y	0.5	0	0.5	0.5	0
31	10	10	10	Y	Y	0	0	0	0	0
32	9	6	8			3	2	3	1	−2
Mean (SD)	9.08 (0.75)	8.63 (1.07)	8.32 (1.20)			0.97 (1.06)	0.74 (0.87)	0.45 (0.76)	0.76 (0.98)	0.32 (1.11)
Agree				31.58%	47.37%					
Max	10	10	10							
Min	7	6	5							

"Bias" scores calculated by subtracting each of video rater 1's scores from each of the live rater's scores, video rater 2 scores from the live rater scores and then video rater 2's scores from video rater 1. If there is no particular tendency for a rater to score high or low, the mean of these values will tend towards 0. If, however, there is a tendency in one direction positive or negative values will be returned.

participants also use different strategies such as asking for help or counting the number of shops using the centre floor plan. Some strategies are easier than others to detect. The correlation between the live and video rater 1 and video rater 2 were (Spearman's rho) 0.78 and 0.69 ($P = 0.001$) respectively. Video raters 1 and 2 correlated 0.64 ($P = 0.001$), with these values being likely reduced by the narrow spread of the scores in this healthy group. Overall, in terms of tasks achieved, therefore there was reasonable agreement between the video and live rating methods but the results suggest that, even when two people are looking at the *same* video clips, discrepancies do occur. This suggests that the reliability of the conventionally used 'live' method may be similarly noisy – a factor that has not been taken into account in previous studies.

3.3. Difference between live video 1 and video 2

Ratings of individual variables between the raters were compared. There was a significant relationship between all the ratings on all the variables of the MET p (two-tailed) <0.05, as seen in Table 5. All variables apart from Mistakes Live and Mistakes Video 2 were

Table 5
Correlations between raters and variables – Live rater, video rater 1 and video rater 2

	Tasks live	Time live	Mistakes live	Tasks video 1	Time video 1	Mistakes video 1	Tasks video 2	Time video 2	Mistakes video 2
Tasks Live									
Time Live	−0.103								
Mistakes Live	−0.367	−0.117							
Tasks Video 1	0.783**	0.147	−0.130						
Time Video 1	−0.062	0.966**	−0.003	0.182					
Mistakes Video 1	−0.424	0.191	0.719**	−0.337	0.302				
Tasks Video 2	0.688**	−0.143	−0.237	0.635**	−0.203	−0.447			
Time Video 2	−0.073	0.967**	−0.037	0.160	0.970**	0.295	−0.163		
Mistakes Video 2	−0.098	0.222	0.503*	0.169	0.283	0.576**	0.124	0.354	

**Correlation is significant at the 0.01 level (2-tailed), *Correlation is significant at the 0.05 level (2-tailed).

Table 6
Correlations between live and video ratings

	Tasks live	Time live	Mistakes live	Tasks video (mean)	Time video (mean)	Mistakes video (mean)
Tasks Live						
Time Live	−0.103					
Mistakes Live	−0.367	−0.117				
Tasks Video (mean)	0.773**	0.029	−0.200			
Time Video (mean)	−0.062	0.975**	−0.012	0.026		
Mistakes Video (mean)	−0.248	0.202	0.683**	−0.059	0.305	

**Correlation is significant at the 0.01 level (2-tailed).

significant to $p < 0.01$. This suggests a strong inter-rater reliability of the task, even when scoring of the task is not live and immediate.

It was also shown that ratings between video rater 1 and video rater 2 were significantly correlated in all variables. As can be seen in Table 5 there was no relationship between any of the other individual variables – scores were not predictive of each other.

3.4. Difference between live and video ratings

Video 1 and Video 2 scores were then collapsed to see if there was a difference overall between live ratings and one single video score – "Video (mean)". Again, from this it was shown that live ratings were significantly correlated with video ratings p (two tailed) <0.01. As demonstrated in Table 6 below, all variables showed a positive correlation in live and video ratings, providing further evidence for the reliability of the method.

Intra-class correlations were used to investigate the inter-rater reliability of the three main variables of the MET – tasks, time and mistakes to see if there was absolute consistency between raters. The correlation between the raters was as follows: $r = 0.825$, $p < 0.001$ for tasks, $r = 0.997$, $p < 0.001$ for time and $r = 0.861$, $p < 0.001$ for mistakes. This shows that there is little disagreement between the raters on the three main variables.

4. Discussion

From this reasonably small sample significant correlation was found on all three key variables by all three raters in live and video scoring of the task and there appears to be strong inter-rater reliability. Discussions between raters on what is meant by each of the "errors" listed were important and proved valuable, as was piloting the video scoring. An example of this was that Video rater 2 was scoring "asking for help" (by a shop assistant) as an "inefficiency" while Video rater 1 noted it as an efficient strategy. From this it was decided to keep a record of the number of times help was asked for but not to score it as an error. Positive results from this study indicate the possibility of using this method of scoring and assessment of the MET in subsequent studies.

The most common error made by participants was the final subtask – telling the rater the time when they had finished – indicating a common difficulty with prospective memory for this group of participants. This the most difficult subtask as it was failed most often. Participants seemed to be so relieved to be finished the task that they forgot this final subtask. It would be interesting to see if this error would be as common if the subtask was to be moved to a different part of the MET – "After ten minutes tell me what time it is". This subtask was at times difficult to detect from the video recordings as some participants turned off the video before they told the time which led to a discrepancy between live and video ratings on this subtask.

Planning time was difficult to accurately measure in this task as some participants stood at the beginning of the task and made a plan as to how they would achieve their subtasks, while others jumped straight into the task and planned as they went along. Although it might have been of interest to get a measure of this it was not possible due to the differing manner of planning that participants used. Miotto and Morris (1998) showed that patients were not found to be any slower than controls indicating that having a score of "planning time" may not contribute over and above other types of score. Other authors have used planning times in their virtual versions of the task and have shown it to make a difference to performance so it could be an issue worth addressing in the future. It would be useful to see if participants were required to plan their route and their time and budget prior to beginning the task as in certain other studies (e.g. Logie, Trawley & Law, 2011; Jovanovski, Zakzanis, Campbell, Erb & Nussbaum, 2012) if an improvement in performance would be seen.

Certain difficulties came from undertaking such a task. A large number of participants declined taking

part due to time constraints or the difficulty of getting to the shopping centre in question. A second difficulty came in carrying out the task in public. Because the task took place in a shopping centre (with permission from the centre's manager) the suspicions of the security staff were raised on many occasions. There were many different staff members working over the testing days. As participants generally followed a similar route, entered many of the same shops, were walking around carrying a clipboard and taking notes over many days and many sessions while being followed around the centre there was often a security guard in turn following the live rater! In order to minimise the potential disruption this might cause – stopping the task to explain to staff what was happening- it was decided not to use the clipboard and to do as much of the preparation and informed consent and payment outside the centre. Shop staff seemed uneasy with customers carrying clipboards but had no problems with customers carrying around pages which potentially had shopping lists written. There were problems too, as mentioned, with the quality of some videos. Two different cameras were used and the second proved much better quality and easier to use. This pen stayed in place better, had better quality audio recording, was less conspicuous, and had better battery life. Most of the previous difficulties that were seen in the first camera were resolved with the second camera. Only one video from the second camera needed to be discarded and this was because the participant was carrying the instruction sheet in front of the camera lens.

With further investigation and validation it would be of interest to see if this assessment measure could also have the potential as a clinical intervention strategy. It could be scored in conjunction with the client in order to identify areas of difficulty and help to devise strategies to make performance of the task more efficient. Knight et al. (2002) suggested using the MET with clients to facilitate the process of goal-setting. It could be useful to have a clear measure of progress over time.

As it is a task that can be used in different settings it would be useful to see if repetition in various settings would lead to improvements in other everyday settings. This would give clients opportunities to practice dealing with unexpected situations, devising efficient strategies etc. in order to achieve small goals set in a real-life but safe situation where they will be able to retrospectively review their own performance. A task such as this also allows clinicians the potential to taper the difficulty and demands of the task by gradually introducing settings that for example have more distracters and more potential for rule breaks or social interaction depending on the level of functioning and goals of the client. Various versions of the task already exist and have been validated, as previously discussed.

Participants in this study were a relatively homogenous group of high functioning older adults. None of the group had any significant everyday organisational problems and although none scored at ceiling for this task it would be of interest to see if scores between live and video ratings still had as strong a relationship when participants experienced more difficulty with the task.

In summary, inter-rater reliability of this task appears strong and lends itself to being used in further investigations with different populations. It provides initial validation for the use of remote scoring of assessments, in this case the MET, in reducing bias and promoting better use of clinician's time as they do not necessarily need to concern themselves with live scoring.

Acknowledgments

We gratefully acknowledge the Grafton Shopping Centre management for allowing us to access and use the centre to carry out our MET. Our thanks go to two visiting students to the unit Veronica Montani and Sarah Griffiths who helped with the video ratings. Without their help this study would not have been possible. Most of all, we thank our participants for the effort they put into this study.

Declaration of interest

The authors report no declarations of interest.

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*, 31-44.

Boyd, T. M., & Sautter, S. W. (1993). Route finding: A measure of everyday executive functioning in the head-injured adult. *Applied Cognitive Psychology*, *7*, 171-181.

Burgess, P. (1996). The Dysexecutive Questionnaire. In B. A. Wilson, N. Alderman, P. W. Burgess, H. Emsley & J. Evans (Eds.), *The Behavioural Assessment of the Dysexecutive Syndrome*. Bury St Edmunds: Thames Valley Test Company.

Burgess, P. W., Veitch, E., de Lacy, A., & Shallice, T. (2000). The cognitive and neuroanatomical correlates of multitasking. *Neuropsychologia*, *38*, 848-863.

Heaton, R. K. (1981). *Wisconsin Card Sorting Test Manual*. Psychological Assessment Resources, Odessa, Florida.

Jovanovski, D., Zakzanis, K., Campbell, Z., Erb, S., & Nussbaum, D. (2012). Development of a novel, ecologically orientated virtual reality measure of executive function: The multitasking in the city test. *Applied Neuropsychology: Adult*, 0, 1-12.

Knight, C., Alderman, N., & Burgess, P. W. (2002). Development of simplified version of the multiple errands test for use in hospital settings. *Neuropsychological Rehabilitation*, *12*(3), 231-255.

Logie, R. H., Trawley, S., & Law, A. (2011). Multitasking: Multiple, domain-specific cognitive functions in a virtual environment. *Mem Cogn*, *39*, 1561-1574.

Miotto, E. C., & Morris, R. G. (1998). Virtual planning in patients with frontal lobe lesions. *Cortex*, *34*, 639-657.

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