CAPTURE24:

Testing self-report time-use diaries against objective instruments in real time

Jonathan Gershuny, Teresa Harms, Aiden Doherty,

Emma Thomas, Karen Milton, Paul Kelly, Charlie Foster

University of Oxford

October 2017

Abstract

This study provides a new test of time-use diary methodology, comparing diaries with a pair of objective criterion measures: wearable cameras and accelerometers.

A volunteer sample of respondents (n=148) completed conventional self-report paper time-use diaries using the standard UK Harmonised European Time Use Study (HETUS) instrument (Eurostat 2009). On the diary day, respondents wore a camera that continuously recorded images of their activities during waking hours (approx. 1500-2000 images/day) and also an accelerometer that recorded their physical activity (PA) continuously throughout the 24-hour period covered by the diary. Of the initial 148 participants recruited, 131 returned usable diary and camera records, of whom 124 also provided a usable whole-day accelerometer record.

The comparison of the diary data with the camera and accelerometer records strongly supports using diary methodology at both the aggregate (sample) and individual levels and provides evidence that time-use data may be a preferable alternative to PA Questionnaires (PAQs) for providing population-level estimates of physical activity energy expenditure (PAEE). It implies new opportunities for calibrating metabolic equivalent of task (MET) attributions to activities, using large scale time-use diary studies deployed for samples representative of national populations.

1. Introduction

1.1 Background

Time-use diary methods are used for a range of research purposes in the social sciences. Economists use diary data to estimate extended National Product measures, including the value of unpaid work (Goldschmidt-Clermont and Pagnossin-Aligisakis 1999). Sociologists employ them to investigate parenting practices (Craig and Mullan 2011), sociability (Voorpostel, van der Lippe and Gershuny 2009) and the division of domestic labour (Sullivan 2000). Whilst diaries are used as a data collection method by some public and population health researchers (e.g. Brunner, Juneja and Marmot 2001), they are not routinely employed to estimate the extent and distribution of time devoted to physical activity (PA) across large populations. Rather, the convention has been to use various forms of physical activity questionnaires (PAQ) that include a battery of items asking respondents to recall the number of times they participated in specific activities over a specified period (last week/month). One of the most routinely used PAQs is the International Physical Activity Questionnaire (IPAQ), or its Short Form (IPAQ-SF).

1.2 Objectives

This paper reports the results of the CAPTURE-24[i] project. The first objective is a test of self-report time-use diary reliability against objective criterion measures. The validity of the daily diary account (as long as this is collected reasonably soon after the events) not actually in doubt, since diarists provide a direct representation of their own activities to the researcher. Nor is the reliability of the camera and accelerometer evidence uncertain, as both instruments record respondents' activities in continuous real time. In this study, they are deployed as *criterion variables*—variables with self-evident reliability, but less secure validity—as straightforward means of checking the timing and duration of the activities recorded by respondents in their self-report time-use diaries.

The second and more specific objective is to argue that time-use diaries are more appropriate than PAQs for many public health research purposes. The PAQ approach has several shortcomings. First, the responses to these types of questions are known to be seriously biased in directions determined both by respondents' perceptions of social desirability (Bernstein Chadha and Montjoy 2001; Shepherd 2003) and by their attempts to enact particular sorts of normatively sanctioned identities (Brenner and DeLammater 2014). Lee, Macfarlane, Lam and Stewart (2011) carried out a systematic review of the validity of the IPAQ-SF and reported that it typically overestimated PA measured by an objective criterion by an average of 84 percent. They concluded that evidence supporting IPAQ-SF as an indicator of relative or absolute PA is weak.

The immediate precursor to the current project was Paul Kelly's doctoral thesis (Kelly 2013, reported in Kelly, Doherty, Mizdrak, Marshall and Kerr 2014), which compared travel behaviour recorded by participants (n=69) wearing an automated SenseCam wearable camera with their registrations in a UK National Travel Survey-type trip log for the same day. The CAPTURE-24 project is the first full-scale attempt to test the accuracy of *continuous* diary records against objective and comprehensive measures (using passive data collection devices) of daily activity recorded in real time.

2. Literature review

There is a surprisingly long history of methodological research into time-use diary reliability studies—most of which relied on the convergence of multiple non-criterion-variable type time-use

estimation methods. The earliest direct test using a real-time activity record as an objective criterion variable involved a video camera placed on top of a television set in 20 US households (Bechtel, Achepohl and Akers 1972). The 'objective' measure of television viewing was obtained by registering the presence of household members sitting in front of the television screen while the set was switched on. This corresponded well with the record of television viewing found in the Cross-National Comparative Time Use Study (Szalai 1972) using general purpose time-use diaries kept by household members over the same period.

Robinson and Godbey (1997) having reviewed a number of previous examples of this type of methodological research (e.g. Robinson 1985, Juster 1985, Hill 1985, Presser and Stinson 1996) concluded that additional controlled studies needed to be undertaken to extend and refine the estimates. Subsequent, methodologically sophisticated approaches to non-criterion-based tests (e.g. Kan and Pudney 2008) reiterate the view that diary approaches can be regarded as a 'gold standard'. In their review, Brenner and DeLamater (2016) report no definitive progress in establishing validity or reliability on grounds other than *a priori*. Without an adequate criterion variable, deductive arguments are mere speculations.

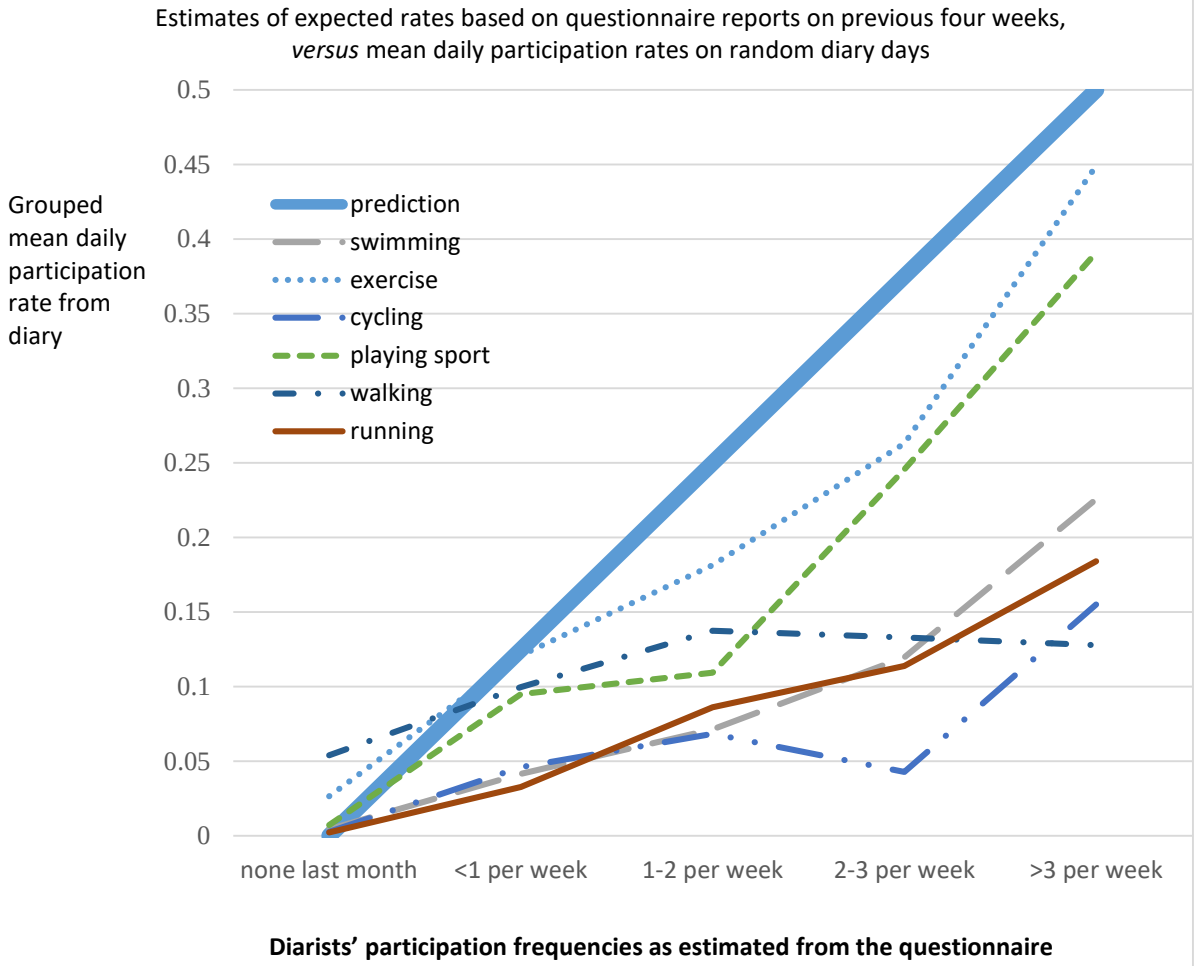
2.1 Estimating PA: Time-use data versus PAQs

Figure 1 (an updated version of Gershuny 2012: 258) drawn from the 2014–15 UK Time-Use Survey (UK TUS) (Gershuny and Sullivan 2017) shows the relationship between the reported rates of PA participation from the questionnaire completed by respondents in the UK TUS, and the participation rates that emerge from their randomly selected diary days (weighted to give an equal representation of days of the week)—a convergent reliability test[ii].

Assuming that past participation rates indicate future participation probabilities, we suggest that any respondent who reported, say 14 or more instances of participation in the past month (i.e. more than 3 per week) would be expected to have a roughly >0.5 probability of participation on a randomly chosen day (re-weighted, as in the previous paragraph). This sort of reasoning gives us the 'predicted participation' line. Diary evidence on participation in walking, cycling, running and swimming provide participation rates of between 0.13 and 0.22 for this group.

About 5% of those who report no walking and 2% of those who report no purposive exercise the previous month show some participation on the randomly chosen day, but with these two exceptions, *all* of the diary participation rates are substantially below what would be expected from the questionnaire answers. The average slope of the swimming, exercise, cycling, sport, walking and running lines is about half-way between the x-axis and the prediction line, which corresponds well with Brenner and DeLammeter's (2014) 'double the actual' estimation and the results from Macfarlane et al. (2011).

Figure 1: Actual vs predicted daily participation (UK 2015 data, our calculations)



Another serious shortcoming is the constrained range of coverage of most PAQ batteries. All daily activities involve some level of physical activity energy expenditure (PAEE), but the PAQ items only cover a limited subset of pre-specified activities. Some respondents' main source of PAEE may be outside the range covered by the PAQ. For example, incidental daily moderate-to-vigorous activities (e.g. caring for babies and toddlers, home renovation, gardening) are not be captured adequately by PAQ items. Someone commuting to work might forget to include running for the bus. By contrast, respondents' detailed 'own words' diary descriptions provide an even coverage across all daily activities resulting in a better-balanced estimation of the extent of different types of PA, although not their intensity.

These two issues with the PAQ approach, together with the centrality of PA measurement to the understanding of obesity, diabetes, cardiovascular disease and cancer (e.g. I-Min Lee, Shiroma, Lobelo, Puska, Blair and Katzmarzyk 2012) provide, in addition to the many social science applications mentioned above, a strong public health-based motivation for the reliability evaluation enabled by the CAPTURE-24 project.

3. Study design and methods

3.1 Ethical considerations

This study received ethical approval from University of Oxford (Inter-Divisional Research Ethics Committee (IDREC) reference number: SSD/CUREC1A/13-262). The study investigators followed the comprehensive ethical framework on appropriate ethical protocols for conducting research with wearable cameras (Kelly et al. 2013). Participants signed a consent form after a member of the research team had fully explained the study requirements. Investigators recommended that participants check in advance that friends, family, and co-workers understood the nature of the study and were happy for them to take part and were also advised of places where wearing the camera may not be appropriate (e.g. changing rooms, banks and schools). All of the cameras were encrypted and did not record sound, voices, or conversations. Before the 'reconstruction' interview, participants were invited to view the images (in private) and to delete all unwanted images without giving a reason. Participants were not allowed to keep any copies of the images.

3.2 Sample and setting

The volunteer sample was drawn from the UK county of Oxfordshire. The research team invited participants via professional networks, free online advertisements, posters, social and sport clubs, word of mouth from other participants, and emails to an authorised list of willing research volunteers provided by a market research agency. Every effort was made to recruit a representative sample across sex, age (18 years and over) and educational level (Table 1). The original sample of 148 participants returned 124 complete diary, camera and accelerometer records, and 131 diary/camera pairs.

Table 1:
Age, sex and educational composition of achieved diary-camera sample

	men	educational level		
	all	missing	below University	University
young adult (18-39)	32	1	8	23
middle aged (40-59)	7		2	5
older (60+)	14	1	7	6
	53	2	17	34
	women			
young adult (18-39)	42		5	37
middle aged (40-59)	27		8	19
older (60+)	9		3	6
	78		16	62

3.3 Design

The study design and associated protocols were refined based on the pilot study findings (n=14) (Kelly et al. 2015). Participants met with a member of the research team before and after the data collection day. The purpose of the initial meeting was to explain the project purpose, gain written informed consent, complete a short demographic questionnaire (including self-reported height and weight to calculate body mass index (BMI)) and receive the three instruments (diary, camera and

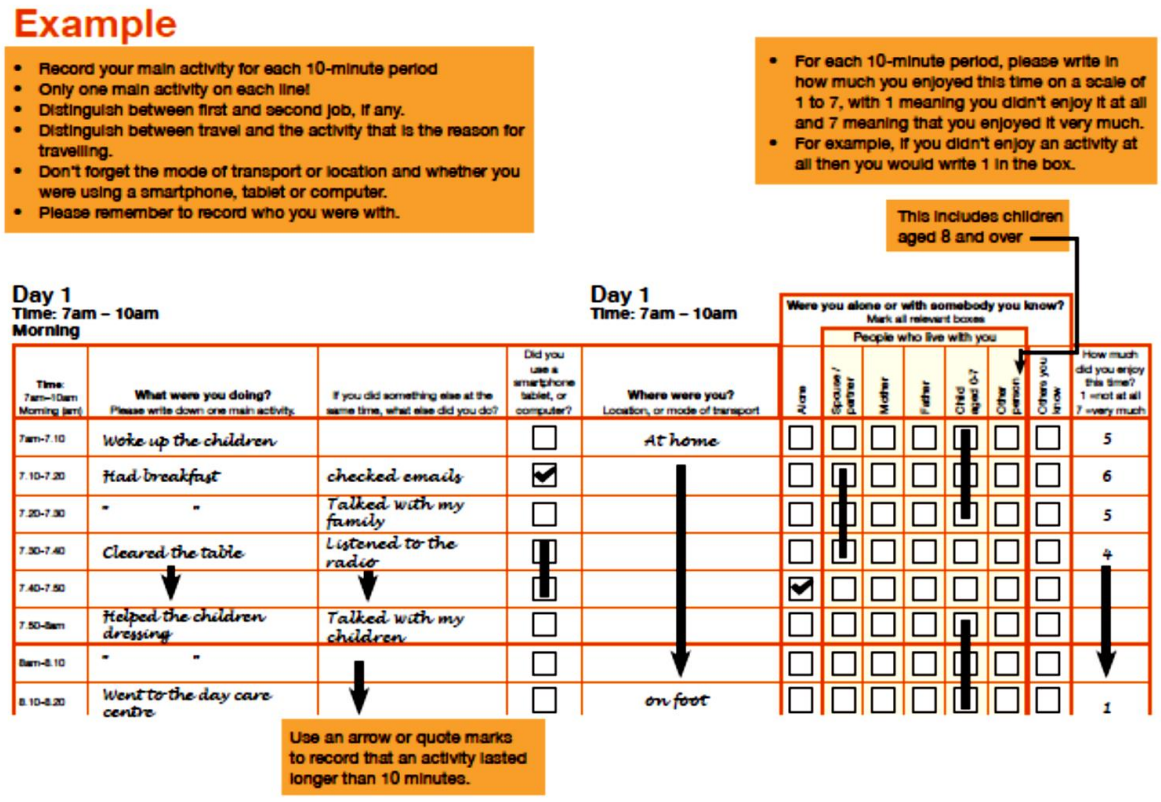
accelerometer) and instructions on how to use them. On the data collection day, participants completed a self-report time-use diary and wore the two passive data collection devices (camera and accelerometer). Shortly after the data collection day, participants met with a researcher for a post-data collection 'reconstruction interview' and to report their experience of wearing the devices and completing the time-use diary. Participants received a £20 High Street voucher after completing the interview.

3.4 Instruments, devices and interview

3.4.1 Time-Use Diary

The diary used in the study was the same as those from the 2014–15 UK TUS, which was the UK version of the European Harmonised European Time Use Study (HETUS) (Eurostat 2009). The diary starts at 4:00 am and covers 24-hours, in 10-minute intervals, with three hours on each page (Figure 2). Participants completed the diary in their own words across six fields or 'domains': primary activity, secondary activity, co-presence, location or travel mode, technology use, and enjoyment. Typically, a one-day diary takes about 20 minutes to complete.

Figure 2: Example page of the UK HETUS self-report time-use diary



3.4.2 Autographer Wearable Camera

The Autographer wearable camera (Figure 3) was developed by the Oxford Metrics Group (OMG). The first model, the Vicon Revue, was followed by the Microsoft SenseCam which have been evaluated in several papers (e.g. Doherty et al. 2013). Participants wore the Autographer (on a lanyard or clipped to their clothing) for as long as possible during their waking hours—generally after showering in the morning and until preparing for bed in the evening. The camera captured images

automatically at 20 to 30-second intervals (medium capture rate) from the wearer's point of view, but no sound was recorded. A privacy lens allowed participants to halt image recording temporarily

On a typical day, the camera captures 1500-2,000 images and also records ambient temperature and light levels. The average 16-hour battery life is sufficient to cover waking hours for most participants. It is not waterproof so participants were asked not to wear the camera if they were engaged in contact or water-based sports. Figure 2 shows examples of typical images depicting everyday life (e.g. exercising, working at a computer, eating, socialising and shopping), illustrating the image quality and resolution.

Figure 3: The Autographer camera and examples of typical images of everyday activities

The camera functions best in good lighting conditions (i.e. daytime and indoors with sufficient lighting). Travelling after dark (particularly in winter) can result in unclear or poor quality images. Occasionally, participants' clothing or hair can obscure the lens, or data may be lost when the camera is turned off for various reasons (e.g. for privacy or unintentionally).

3.4.3 Axivity AX3 band accelerometer

The AX3, first released in 2012, is a continuous logging accelerometer designed for a range of applications including PA monitoring and classification, motion analysis and medical research (Doherty et al. 2017). The AX3 is compliant with the OpenMovement data format, has sufficient memory for 14 days continuous logging at 100Hz (512MB), is waterproof to 1.5 meters and includes temperature and light sensors. It has an in-built, accurate clock and calendar which provide the time stamp for the recorded acceleration data (axivity.com/files/resources/AX3). The AX3 has configurable sample rates, adjustable sensitivity and a low power mode. The sample rate of 400 Hz gives a battery life of 5 days. The AX3 can be set to different sensitivity levels for specific research applications.

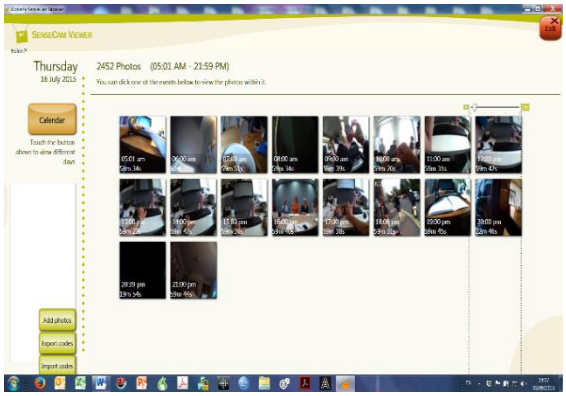
Participants wore the accelerometer (Figure 4) for at least 24-hours on their dominant hand (wrist), although many wore it for a day before and after the diary day, which provided an additional two days of sleep data. As the AX3 has a long battery life and is robust and water-proof, participants were able to wear it while working, travelling, taking a bath or shower, sleeping and playing all types of sport.

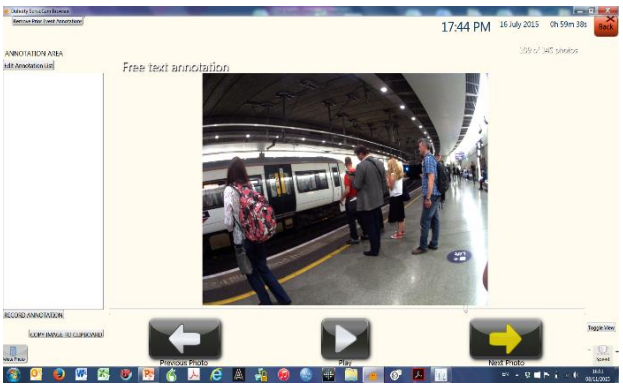
Figure 4: The Axivity AX3 band accelerometer (activity watch)

3.4.4 'Reconstruction' interview

Shortly after the data collection period (maximum four days), participants viewed the camera images in a face-to-face 'reconstruction' interview, which took about 60 minutes. This process is similar to a 'yesterday' diary, but attaining higher validity due to the image prompts (e.g. Cowburn et al. 2015). Before the interview, the investigator downloaded the images into a bespoke browser (Doherty, Moulin and Smeaton 2011) (Figure 5) and invited the participant to view and delete (in private) any unwanted images. Using the images as prompts, participants described their day while the interviewer kept detailed notes to assist with the coding process.

Figure 5: The browser images in thumbnail (a, left) and single-image (b, right) modes





4 Data coding

The reliability test focus makes it essential to code the diary and image data independently. Resource constraints allowed only a single coder, so to avoid contamination, the diary and image coding were carried out separately, approximately four months apart (first diaries, then images). The large number of respondents, combined with the anonymity of the records, meant that the coder had no means of connecting particular diaries with the corresponding image files.

4.1 Time-use diary coding

The HETUS diary instrument uses 10-minute intervals ('time slots'). A time-use *episode* is a sequence of time slots through which there is no change in any of the six substantive domains. The 10-minute interval makes it difficult for diarists to record short-duration (e.g. going to the toilet, checking text

messages) or momentary activities (e.g. taking medication, using an ATM) occupying less than 5 minutes, so activities of less than 5 minutes' duration sometimes fail to appear (though in such cases they may appear in the secondary activity field). The final coded diary data file comprises, for each study participant, a sequence of episodes of varying lengths, starting at 4am, with a total duration of 1440 minutes (Eurostat 2009).

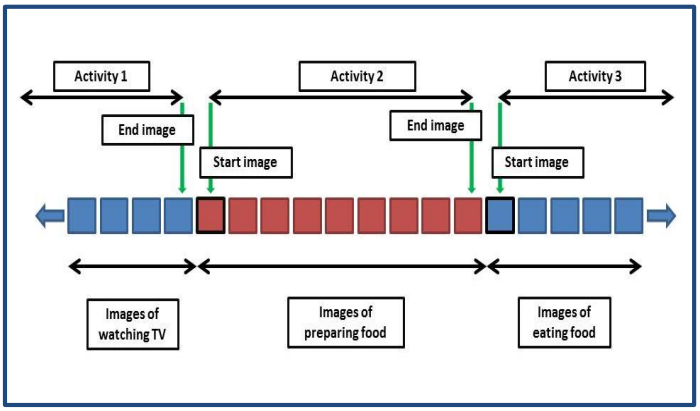
The HETUS activity coding system is hierarchical, to the 3-digit level[iii]. Primary and (up to three simultaneous) secondary activities are coded using the UK version of the standard HETUS activity classification, with just under 300 different activities. Coders categorise the main and secondary activities, location/mode of transport and other domains, and determine the start and end time of these episodes.

4.2 Camera image coding

We applied the same coding procedures to the raw camera images and the diaries, with two exceptions. First, the recording intervals were one-minute, giving the image file a finer granularity than the diary. Second, the enjoyment domain was not used. For the purposes of the diary versus camera comparisons discussed in the following sections, the one-minute intervals in the image files were concatenated to 10-minute diary intervals to allow analysis.

The interview notes were essential to the coding process. Most participants had a few black or unclear images from using the privacy lens cover, inadvertently covering it with clothing or being in low-light conditions, so the interviewer needed to identify what the respondent was doing when this occurred. The main reasons for covering the lens or turning the camera off were showering, reading confidential documents on the computer, attending medical appointments and collecting children from school. The interview notes also allowed the coder to include additional domain information such as secondary activities, location and the presence of others.

Figure 6: The SOP for image coding



We developed a standard operating procedure (SOP, Figure 6) for the image coding to aid replicability. Activities were identified as episodes and assigned a HETUS code if they continued for 3+ images with no 'breaks' of more than 2 images. Activities that lasted fewer than 3 images were grouped with the activity immediately preceding them. For example: 10 images of watching TV → 2 frames of food preparation → 25 frames of watching TV would be coded as a single activity *watching TV*. If the food preparation lasted 3+ images, it would be coded as *preparing food* with *watching TV* on either side (Figure 5 example). One of the limitations of the protocol is that it cannot assign either

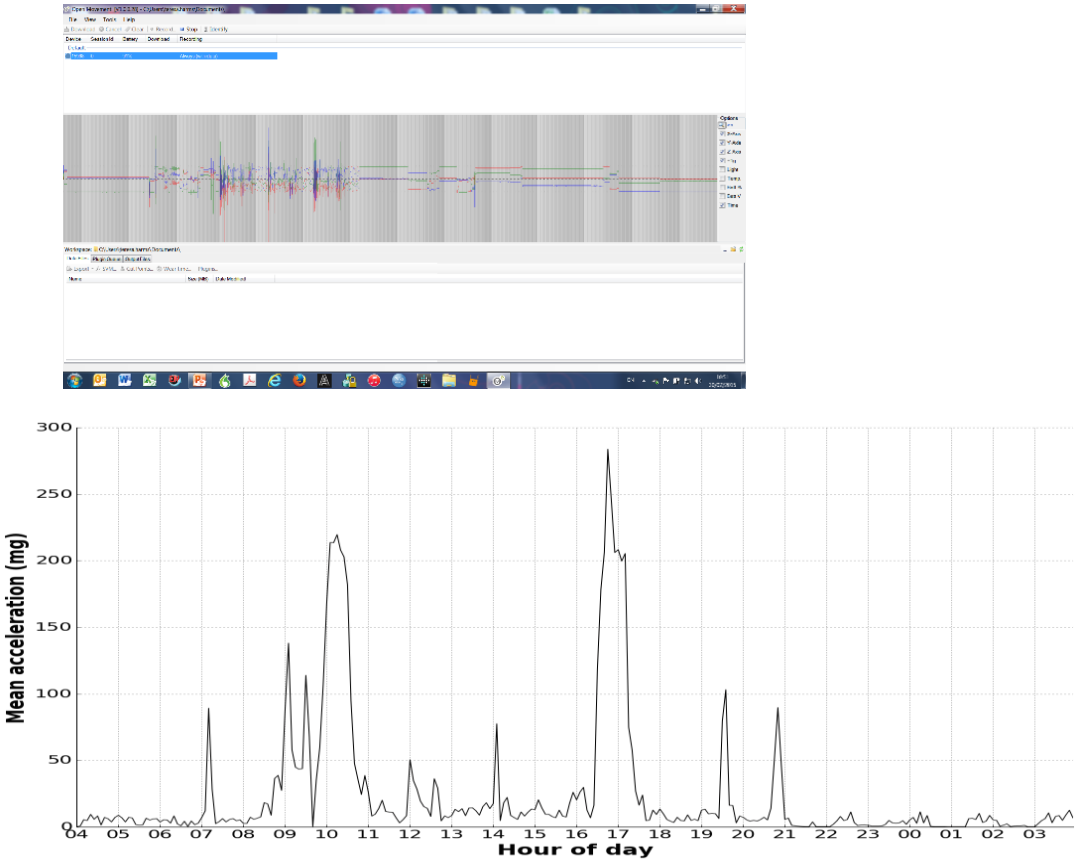
preparing food or *watching TV* as primary or secondary activities unless it was recorded this in the interview notes.

4.3 Accelerometer data extraction

For the accelerometer data processing, we followed procedures used by the UK Biobank accelerometer data processing expert group, including device calibration to local gravity, and resampling to 100Hz. We calculated the sample level Euclidean norm of the acceleration in x/y/z axes, and removed machine noise using a fourth order Butterworth low pass filter with a cut-off frequency of 20Hz. In order to extract the activity-related component of the acceleration signal, we removed one gravitational unit from the vector magnitude, with remaining negative values truncated to zero. Device non-wear time was automatically identified as consecutive stationary episodes lasting for at least 60 minutes.

To describe physical activity intensity, we aggregated the sample level data into five-minute episodes for summary data analysis, maintaining the average vector magnitude value over the epoch (in milli-gravity units). Accelerometer measures that represent total activity volume, such as average vector magnitude, have been recommended as appropriate measures of PAEE. Each signal was summed over a minute (Figure 7).

Figure 7: Raw (above) and converted (below) data from the AX3 band accelerometer



5. Data analysis and results

5.1. Aggregate comparison of diary and camera records

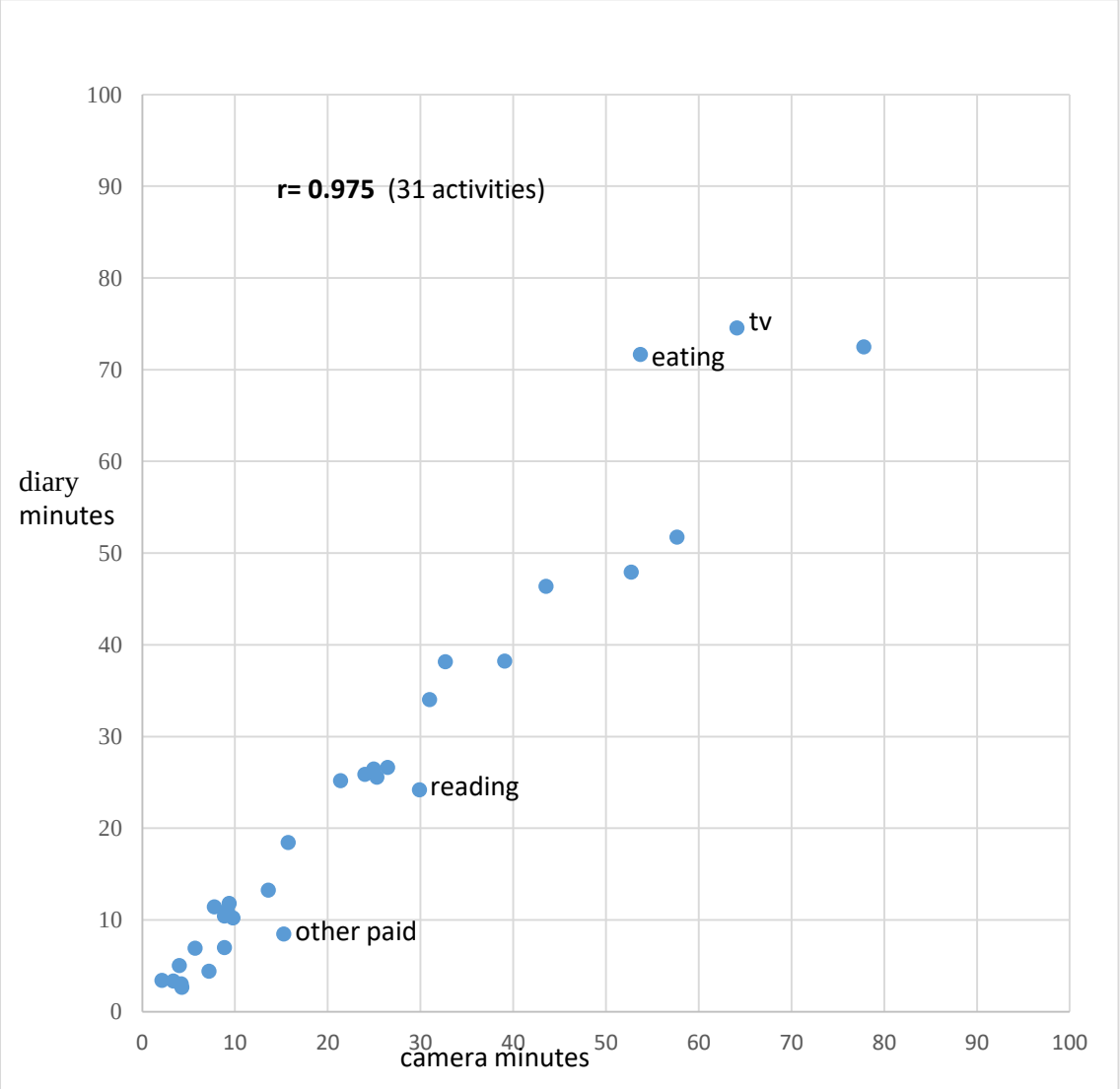
The 33 activities listed in Table 2 comprise activities coded to the 2-digit level of the UK HETUS activity lexicon, together with some amalgamation of activities associated with very small time expenditures. The aggregate mean times in coded activities from the camera data and the self-report time-use diaries are, in general, rather similar. Table 2 shows substantial differences in just three activity categories out of the total of 33 activities: *eating*, *reading* and watching *television*.

Table 2: Mean daily time in 33 activities

	camera		diary	
(131 cases)	mean	*std error*	mean	*std error*
Sleep	497.5	*8.1*	490.0	*7.6*
Eating	53.7	*3.3*	71.7	*4.7*
other personal care	57.6	*2.9*	51.8	*2.6*
main job	183.4	*19.0*	178.5	*19.0*
other paid-work-related	15.3	*4.6*	8.5	*2.6*
school or university	43.5	*10.0*	46.4	*10.6*
food management	52.7	*3.8*	47.9	*3.7*
household upkeep	32.7	*3.0*	38.2	*4.5*
make, dare for textiles	13.6	*2.9*	13.3	*2.7*
gardening and pet care	15.7	*4.1*	18.5	*4.6*
construction & repairs	3.4	*2.4*	3.4	*3.0*
shopping and services	26.5	*3.2*	26.6	*3.1*
household management	8.9	*1.8*	7.0	*1.9*
Childcare	24.0	*4.9*	25.9	*4.8*
organisational work	8.9	*4.0*	10.5	*4.4*
help to other households	4.2	*2.1*	3.1	*1.7*
participatory activities	4.3	*1.4*	2.7	*1.1*
social and entertainment	77.8	*8.2*	72.5	*8.4*
entertainment & culture	9.2	*3.4*	11.0	*3.7*
resting & time out	7.8	*1.5*	11.5	*2.4*
physical exercise	21.4	*3.4*	25.2	*4.0*
arts and hobbies	9.8	*2.9*	10.2	*3.0*
Computing	39.1	*5.4*	38.2	*5.0*
Games	7.2	*2.6*	4.4	*2.2*
Reading	29.9	*4.5*	24.2	*3.7*
Television	64.1	*6.2*	74.6	*8.0*
radio and recordings	2.1	*1.1*	3.4	*1.3*
work travel	31.0	*4.1*	34.1	*4.5*
education travel	9.4	*2.1*	11.8	*2.5*
unpaid work travel	25.3	*2.9*	25.6	*2.9*
civic travel	4.0	*1.2*	5.0	*1.5*
leisure travel	25.0	*4.3*	26.5	*4.6*
exercise travel	5.7	*1.3*	7.0	*1.5*
unclassified time	25.6		11.1	

Figure 8 plots the 31 activity categories with durations less than 100 minutes (excluding *sleep* and *paid work*, both with long durations, as they would distort the view, as well as give a correlation coefficient indistinguishable from unity). It shows a very strong association between the two measures as estimators of time use at the aggregate level. If we just take the 31 two-digit activities as cases, we arrive at a correlation coefficient, between the diary and camera estimates, of .975, which is a remarkably high level of association between a self-report estimate and a criterion measure. Compare, for example, this nearly 45° plot with the divergence between the diary and questionnaire predictions in Figure 1.

Figure 8: 31 activities <100 minutes



5.2 Individual-level comparisons of diary and camera reports

The similarity between the aggregate means of this quite detailed activity list is not entirely surprising. For example, it may be generated by perfect recall of the *sequence* of yesterday's activities, combined with a random error term in the recall of the *start/finish* time of each element in the activity sequence. The errors are self-cancelling across the sample, so as to produce the

unbiased mean estimates seen in Figure 8. Next, we turn from the comparison of *aggregate* mean time in activities across the sample, to consider the patterns of difference between the diary and camera estimates of total time in the activity at an *individual* level (i.e. moving from between-individual to within-individual comparisons).

The main issue, for the present purpose of assessing the reliability of the diary record, is whether we can find statistically significant differences between diary-based estimates of the individual's total time in various activity categories, and the estimates derived from the (criterion) camera record. The t-tests in Table 3 show strongly significant differences only in the case of time devoted to *eating* and more weakly significant results for *other personal care*, *food management*, *reading* and *school travel*.

Table 3 General mutual confirmation of diary and camera records

		* p<0.05 ** p<0.005	*** p<0.0005	
	T-test		**Correlation**	
	(2-tail)	Sig difference	R	Significance
Sleep	0.086		0.847	***
Eating	0.000	***	0.550	***
personal care	0.023	*	0.572	***
main job	0.121		0.986	***
other paid-work-related	0.123		0.353	***
school or university	0.464		0.928	***
food management	0.029	*	0.832	***
household upkeep	0.086		0.706	***
make, care for textiles	0.867		0.784	***
gardening and pet care	0.071		0.946	***
construction & repairs	1.000		0.993	***
shopping and services	0.917		0.831	***
household management	0.283		0.563	***
Childcare	0.396		0.899	***
organisational work	0.305		0.937	***
help to other households	0.506		0.607	***
participatory activities	0.150		0.631	***
social and entertainment	0.344		0.774	***
entertainment & culture	0.131		0.945	***
resting & time out	0.136		0.274	***
physical exercise	0.124		0.789	***
arts and hobbies	0.685		0.929	***
Computing	0.856		0.614	***
Games	0.199		0.604	***
Reading	0.039	*	0.799	***
Television	0.088		0.655	***
radio and recordings	0.389		0.232	**
work travel	0.151		0.882	***
education travel	0.034	*	0.884	***
unpaid work travel	0.890		0.762	***
civic travel	0.368		0.670	***
leisure travel	0.375		0.929	***
exercise travel	0.117		0.844	***

Table 3 also provides measures of the covariance of the two measures (i.e. correlation coefficients). The correlation coefficients can provide an estimate of the extent of 'noise' associated with recall errors in the start/finish times of diary activities, although it is not clear what should be considered a 'good' correlation in this context. For example, we might point to *other paid work-related* (mean 15 minutes in the camera record), *resting and time out* (mean 8 minutes) and *listening to radio and recordings* (2 minutes), all with correlations <.5 as disappointing. However, the major time use categories (>60 minutes per day in the diary record) *sleep*, *paid work*, *social activity*, watching *television* all have correlations > .65. Of the 33 activity categories, nine have r >= .9, seven >=.8, and a further five have r >=.7.

5.3 Simultaneous activities and the construction of daily narratives

It is not coincidental that the major activity categories of *eating*, watching *television* and *reading* show the most substantial differences at both aggregate (sample) and individual (case) levels as these activities are the most likely to occur simultaneously with other activities.

Most participants would be accustomed to being asked *What did you do today?* Answering questions such as this trains individuals to construct narratives such as 'arrived home from work, put the kettle on and made tea, then watched television'. These accounts are, in effect, 'streams of behaviour' in different environments, or sequences of activities that can be nested hierarchically (Barker 1963, Barker, 1968, Barker 1978, Harms 2004). From the diarist's perspective, other simultaneous actions (e.g. sipping tea, glancing at the newspaper) may occur *within*, and evidently *secondary to* the main activity of 'watching television'.

All simultaneous activities reported in the diaries and interviews were coded. However, if the respondent did not nominate the primary activity in the reconstruction interview, it was not always self-evident which activities were primary or secondary/simultaneous. In these cases, we made analytical judgements in order to reconstruct the respondent's 'behaviour stream' in a logical sequence. However, our judgements may have differed from the diarist's subjective understanding of the particular activity (hence our reluctance to consider camera evidence of activity as a straightforwardly valid indicator). Interpreting images from the wearer's perspective (i.e. facing outwards) may also may also lead to problems. A respondent eating a meal may turn to chat to her companion, causing the camera to face away from the plate for a few frames. The analyst, for lack of other evidence, may classify this as conversing, even though the respondent would classify the primary activity as eating, with conversing as a secondary activity.

Table 4 Time-reporting hierarchy as seen in the camera record (mins/day)

seen from:	eating	tv	reading
primary only	55	64	30
primary + 1 secondary	108	97	42
primary + 3 secondary activities	115	101	43

We illustrate these problems by considering the full accounts of three activities in the entire camera record. *Eating* as a primary activity occupies 55 minutes in the camera record compared with 74 minutes in the diary. If all the events in which eating is recorded as a secondary activity were

reversed to place eating as the primary, then eating durations would double. Similarly, *watching television*, 75 minutes as a primary activity in the diary but only 64 in the camera, increases by 50% if television viewing events counted as secondary by the camera analyst are recoded as primary. *Reading*, by contrast, is frequently ancillary to other activities. For example, during a meal, a respondent may read the newspaper rather than converse. The newsprint may feature frequently in the images alongside the plate of food, but from the diarist's perspective, eating the meal is the main activity.

5.4 Are there reporting differences by educational levels?

The issue here is not whether there are variances in the detail of activity reported by respondents with different levels of educational attainment, as plainly we expect such differences. Rather, the question we ask is whether there are substantial *differences in the differences* between the camera and diary. Put more directly, we need to establish whether better educated respondents are more likely to under- or over-report particular sorts of activities in their diaries as contrasted with the camera evidence. Table 5 compares the ratios of camera minus diary differences as a percentage of the diary mean estimates of time in the activities. In this analysis, we emphasise activities which occupy a relatively large part of the average day. Activities occupying 30 or fewer minutes per day have a relatively large number of zero-scores, meaning that either the diary or the camera evidence may be absent.

Table 5. Reporting bias from educational level? (activities >=30mins/day in bold)

	camera minus diary mins as % of diary mean mins			camera minus diary mins as % of diary mean mins	
	school	University		school	university
sleep	**1.1**	**1.7**			
eating	**-26.6**	**-24.6**	social, entertainment	38.5	-2.0
other personal care	**8.2**	**12.8**	cinema, theatre etc	-22.7	-9.5
main job	**-0.3**	**3.4**	resting etc	-43.0	-23.4
other paid work-related	137.8	75.0	physical exercise	-36.7	-7.4
school, university	0.5	-7.2	arts, hobbies	-0.8	-9.0
food management	**12.1**	**9.4**	computing	**1.2**	**2.6**
household upkeep	**-7.5**	**-16.8**	games	474.3	6.1
textile care	-1.8	6.2	reading	**21.6**	**24.9**
gardening, pet-care	**-9.8**	**-20.9**	television	**-17.7**	**-12.2**
DIY	190.0	-13.9	radio	-78.9	-10.7
shops, services	**4.7**	**-4.4**	work travel	8.5	-14.1
h/hold management	81.7	12.6	educational travel	-29.3	-17.3
childcare	-15.2	-5.7	unpaid work travel	**0.4**	**-2.1**
organizational work	6.7	-24.0	civic travel	-17.5	-22.1
help other households	121.9	-17.9	*leisure travel*	***-37.8***	***8.2***
participatory activities	**265.0**	**47.3**	exercise travel	-24.1	-15.0

Most of the larger items in Table 5 show reasonable correspondence between the recording patterns of the higher- and the lesser educated respondents, implying low levels of bias. Among these activities, *sleep*, *eating*, *paid work*, *cooking*, *reading* and *watching television*, show similar

patterns of difference between the two sorts of records. *Household upkeep*, *gardening* and *pet-care* show larger differences, although with the same sign on the errors. Only *shopping, social entertainment* and *leisure travel* show large discrepancies in different directions. Among the shorter-mean duration activities other *paid-work-related*, *helping other households* and *playing games* show substantially lower estimates in the diary records relative to the camera estimates among the less well-educated. *Radio listening*, *resting*, *exercise* and *exercise-related travel* show higher levels of under-reporting among the less well- educated respondents.

5.5 Self-similarity analysis of diary and camera records

We now consider similarities in the *overall patterns of time use* produced by the camera and diary pairs in a more holistic way. (We could focus on the similarity of *timings* of daily activity using the sequence analytic techniques discussed by Lesnard (2010) and others, but we reserve this analysis for another paper.) Instead we now consider the *overall daily totals* of time in activities, using the measure invented by Robinson and Converse (1972)[iv], calculating Generalised Euclidean Distances (GEDs) between pairs of records. By considering each of the 33 activity categories as an independent dimension, we can define a 32-dimensional hypotenuse-equivalent, as the square root of the sum of the squared differences between the paired camera and diary estimates of total time in each activity. The resulting 'self-similarity' measure is the GED between the two time-use measures for a single respondent.

We can also calculate a similar GED between each of the 131 diary records and the camera records of each of the *other* 130 respondents, producing 'general similarity' measures. The self- and general-similarity measures together provide a 131*131 matrix of GEDs, each row corresponding to a diary record and each column to a camera record, with the major diagonal elements containing the self-similarity measures, and the off-diagonals the general similarity measures.

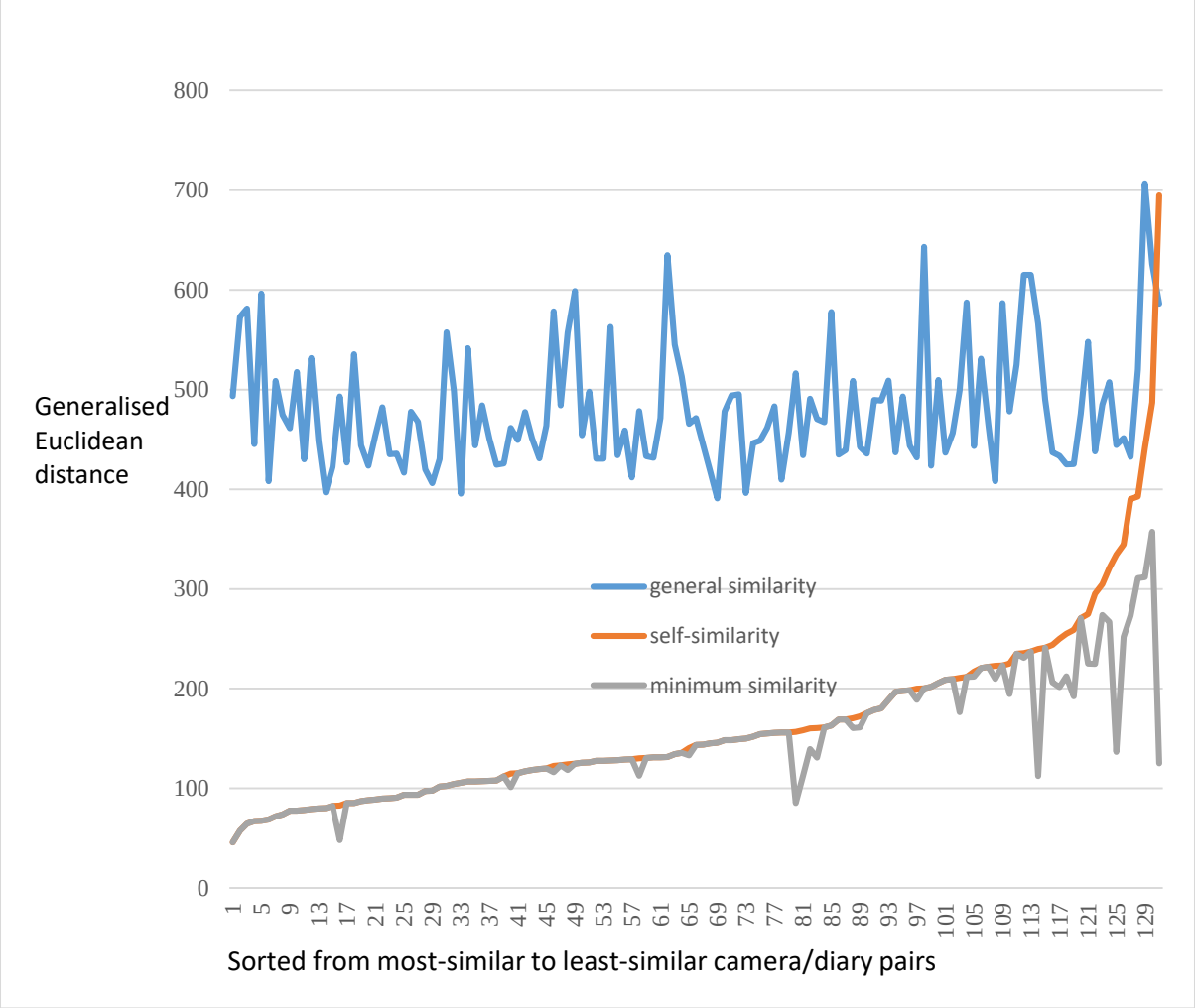
The ratio of the mean of these general similarity measures along a given row of the matrix to the self-similarity measure (the major diagonal cell) provides a goodness-of-fit indicator. We expect, given the extent of interpersonal variation in patterns of daily time use, that the GED between any diary activity pattern and that of the corresponding camera should be smaller than any of the other GEDs between a diary and any of the other camera record; the major diagonal cell should, in general, show the minimum GED on any given row.

Figure 9 reorders the rows and columns of the matrix in ascending order of the 131 self-similarity scores and for each case, plots the mean of the general similarity indicator, the self-similarity indicator, and the minimum GED for the appropriate row of the matrix. The GED scores for each subject, roughly speaking, represent the sum of the deviations between the 33 time use totals from camera/diary pairs; a GED of 100 units represents an average 3-minute deviation for the 33 pairs, 200 represents a 6-minute average deviation, and so on. With the exception of the single worst case, the self-similarity distance is smaller than the mean of the general similarity scores. Likewise, the self-similarity distance for most of the first 100 or so of the re-ordered cases is also the minimum GED. Beyond this point we find an increasing number of cases where the overall time use pattern in the diary record is more similar to someone else's camera record than to the diarist's own record.

As already noted, there are two likely explanations for the differences between the camera and diary pairs. The first is simply poor diary-keeping, which emphasises the importance of checking diaries for missing data upon collection. The second is the difference between the respondent's own

recorded sequence of primary activities from the more complex multiple-simultaneous-activity reality of the camera record, and the coder's decisions. Although beyond the scope of this paper, this can be tested by observing the effects of re-ordering the multiple simultaneous activities recorded by coders in the camera records (for example in Table 4).

Figure 9: Comparison of similarity of diary/camera pairs and distance of diaries to means of all other camera records



There are several documented indicators for diary quality (e.g. Fisher et al. 2015; Glorieux and Minnen 2009). These include: (1) range of coverage in the daily record (i.e. its inclusion of necessarily daily activity such as eating or sleeping); (2) the frequency of mentions of secondary or higher-order simultaneous activities; (3) the amounts of missing time during the day and; (4) the number of separate activities recorded in the diary. In this analysis, we deploy the latter two indicators. Removing 'low quality' diaries (defined as those with more than 60 minutes missing/unallocated time during the diary day and with fewer than 25 diary episodes) leaves 100 'high quality' diary records of the 131 total. Of these, 90 have self-similarity scores of no more than 15 units (i.e. average deviations of less than 30 seconds above the minimum for their case).

5.5 Aggregate comparisons of diary, camera and accelerometer measures

Table 6 groups the 33 two-digit activities into seven broad categories and compares the PA levels (accelerometer records in mg/minute) associated with each. The upper two panels of the table refer

to the camera records. On the right are the means and standard deviations for all participants who completed diaries and on the left the 'high quality' diaries. Only a subset (n=124) of the camera and diary sample returned usable accelerometer data. In order to maintain adequate numbers, we used a slightly less stringent criterion for diary quality, classing all those with fewer than 70 minutes missing as 'good' diaries. The lower two panels provide equivalent measures comparing the diary to the accelerometer records.

Table 6. Comparison of accelerometer means for summary activities, by camera and diary

Mean mg per minute						
(case=diary day)	'good' diaries only			all diaries		
	Camera records					
	Mean	Std Dev	N	Mean	Std Dev	N
Sleep	19.7	15.0	94	19.5	14.7	100
Leisure	35.6	16.2	100	35.4	16.7	118
paid work	41.7	29.7	63	41.3	29.1	74
eating, personal care	69.3	25.7	100	68.5	25.7	117
unpaid work	64.7	25.5	99	65.0	26.7	118
Travel	89.8	43.3	99	89.2	43.2	118
Exercise	158.3	142.9	38	173.6	159.9	43
	Time-use diary records					
	Mean	Std Dev	N	Mean	Std Dev	N
sleep	19.5	14.7	99	19.4	14.7	100
leisure	36.2	18.9	104	36.0	19.2	114
paid work	40.8	27.9	66	40.2	27.5	77
eating, personal care	57.2	20.4	102	56.6	20.8	113
unpaid work	70.3	26.8	101	68.4	27.1	115
Travel	84.6	39.2	105	83.5	38.5	119
exercise	161.8	140.4	38	172.8	154.5	41

Two findings emerge with some clarity from Table 6. The first is that both the camera and the diary records show the expected differences in PA between the broad types of activity. For example, in all four quadrants of the table we find a roughly eightfold difference in PA between the *sleeping* and *exercise* categories. In particular, the same differentials emerge from the camera and diary records.

The second finding, with a single exception, is that there are insubstantial differences between the whole sample and the 'high quality' diary columns. The exception is *exercise* (e.g. sports, walking), where diaries from the whole sample report higher levels of PA than the 'high quality' diaries: 174 mg/min versus 158 mg/min for the camera records, 173 mg/min versus 162 mg/min for the self-report diary. The standard deviations of these means are large, which indicates that these differences are not statistically significant.

Although the precise mechanism is not clear, in both cases the less densely-recorded diary and camera sequences reveal somewhat more exercise. Perhaps, in these cases, activities such as running for a bus or taking the stairs (which might otherwise be classified in a leisure, paid work or travel category) were instead placed in one of the subcategories of exercises, therefore slightly

reducing the 'all participants' mean PA in the former categories and substantially increasing it in the latter.

Table 7 Accelerometer means by 2-digit activity categories, ordered by camera scores

(124 cases)	Camera	Diary		Camera	Diary
sleep	19.1	18.7			
radio and recordings	14.5	29.2	childcare	59.7	60.5
television	27.3	30.2	food management	67.7	64.7
reading	30.0	30.4	help to other households	65.6	64.9
participatory activities	38.8	31.4	leisure travel	75.5	66.6
school or university	31.1	31.8	other personal care	73.2	71.9
computing	35.0	32.6	construction & repairs	69.3	73.0
games	26.4	35.1	household upkeep	82.1	74.6
household management	42.7	41.0	make, care for textiles	92.8	80.2
resting & time out	47.6	42.5	shopping and services	76.2	80.7
main job	43.9	43.4	gardening and pet care	91.1	84.4
arts and hobbies	43.1	44.0	work travel	92.8	86.7
social and entertainment	45.1	44.6	civic travel	91.1	88.5
eating	41.7	44.6	education travel	91.2	91.8
entertainment & culture	35.5	48.1	unpaid work travel	101.2	92.3
other paid-work-related	56.6	51.2	exercise travel	93.1	92.7
organisational work	44.9	51.8	physical exercise	176.9	172.8

Table 7 compares the mean accelerometer scores, broken down both by the camera and the diary classification of each activity for the more detailed 2-digit activity classification. The rows of the table are placed in ascending order of the diary-based accelerometer scores. The ordering would differ only slightly—activities moving up or down by no more than a single rank—if it were reordered according to the equivalent camera coding. There is a correlation of .98 between the scores derived from the camera- and diary-based coding. (We excluded scores for exercise from our calculation of this correlation because, as distinct outliers, they would bias the estimate upwards.[v])

5.6 Individual-level comparisons of diary, camera and accelerometer measures

Just as we did for the 2-way diary and camera analysis, we now turn from sample means to an individual-level analysis. We start with a simple OLS regression of the camera and diary-based classification of each 10 minute time slot through the 1440 minutes of the day, on the mean accelerometer score for that timeslot. The timeslot is the 'case', yielding a potential dataset of 17,856 (i.e. 124*144) cases for both the diary and camera records, although missing data reduced this total to 16,846 cases for the records that have valid camera, diary and accelerometer measures for the same time slot.

The simple OLS approach to this is a 'dummy variable regression', classifying each time slot-case by a vector of 32 indicator (0/1) variables representing the activity categories, 31 of which are always set to zero. The 33rd 'default' activity category is represented by the case where none of the indicator variables are set to 1. The camera-based regression analysis of the whole dataset produces a

multiple correlation (R) coefficient of 0.493, the diary only slightly less at 0.473[vi]. Considering that much of the variation in PA relates to physiological, demographic and socio-economic variables (BMI, fitness, age, sex, employment status, social class, etc.) that can vary almost-independently of the type of activity, these are reassuringly acceptable levels of association from the perspective of the reliability of the two alternative indicators (i.e. camera and diary) of the type of activity in the timeslot.

However, assessing the reliability of the diary using the camera record as a criterion indicator requires a slightly different approach. It is important to know whether the diary measure is explaining the *same part* of the variation of the accelerometer record as is the camera measure. We modelled this by allocating MET[vii] scores—using the Ainsworth Compendium (Ainsworth et al. 2011) as a reference—to the 3-digit HETUS activity classification. Our process broadly duplicated the work carried out by Tudor-Locke, Washington, Ainsworth and Troiano (2009) who applied this to the American Time Use Study (ATUS) activity lexicon. The raw correlations between the camera- and diary-derived METs scores on one hand, and the accelerometer measure on the other, are 0.518 and 0.500 respectively.

Table 8 provides multiple correlation scores for model 3, which deploys both camera and diary estimated METS to predict accelerometer scores. The relatively small increment of prediction gained by adding the camera METS above the diary METs suggests that both the camera and diary are explaining *the same* components of the variation in the accelerometer record. Adding descriptors of the respondents (e.g. age, sex and educational attainment) improves the model performance, but we reserve further modelling of METs for another article.

Table 8
Diary- and camera-based METs as predictors of accelerometer scores

	model 1		model 2		model 3	
Multiple R	0.500		0.518		0.546	
Multiple R^2	0.250		0.268		0.298	
Mean METs, camera			27.495	***	17.409	***
Mean METs, diary	25.954	***			13.209	***
Constant	-0.287	***	-4.979	***	-10.703	***

6. Discussion

The overall purpose of the CAPTURE-24 project was to test the self-report diary method of capturing time-use information, in a comprehensive way, against records of activity that are sufficiently objective to be considered as *criterion tests*. This is the first occasion, in either the social scientific or the public health literature, that such a test, covering all the activities of daily life, has been carried out.

A prior question to consider briefly is why we do not use the criterion variables *themselves* instead of diary measures. For some research purposes (e.g. dietary analysis) wearable cameras are appropriate (e.g. O'Loughlin 2013), whilst for other topics (e.g. sleep) accelerometers are more suitable (e.g. van Hees et al. 2015). However, for more general purposes that require comprehensive and detailed coding of daily activities, the camera records—requiring both reconstruction interviews

and painstaking re-coding tasks—involve substantial extra costs (i.e. a similarly funded diary study alone might have achieved ten or more times the sample size discussed in this paper). Furthermore, while some activity categories (e.g. sleep, vigorous exercise) can be validly inferred from accelerometer variables, we are at present far from being able to infer the generality of daily activities from physical movement evidence alone.

The sample studied here is in no sense representative of any specific population. Despite our efforts to arrive at a broad base of recruitment of participants, the possibility remains that there is some hidden bias towards unusually accurate diarists. However, our investigation of the relationship of educational levels to reporting provides evidence of a systematic bias from this source.

Holding this issue on one side, we demonstrate that self-report time-use diaries provide a reliable basis for the accurate estimation of time-use patterns, without evidence of undue bias by educational level. By direct inference, we can therefore conclude that when collected from representative samples of respondents, time-use diaries can validly and reasonably reliably represent the time-use of populations. This is an important advance on the previous time-diary evaluation literature, insofar as it relies not on *a priori* reasoning but on comparisons with unimpeachable criterion data.

Our results amplify, on a much broader basis, the conclusions of Kelly *et al* (2014) comparing self-report trip logs to camera records of travel: the self-reports provide generally accurate and unbiased aggregate estimates of means of time in activities, with a random error at the level of individual observations, presumably related to recall error. The CAPTURE-24 study is the first to provide a clear test of the performance of conventional standard time-use diaries against reasonably objective criterion measures and covering the full range of daily activities.

The final observations relate more specifically to methods for estimating PA in the context of public health research. Combining the generally supportive evaluation of the diary against the camera and accelerometer in the two criterion-variable-based assessments, with the poor convergent reliability exhibited in the camera/PAQ comparison illustrated in Figure 1, we conclude that the PA battery is an insufficient and perhaps inappropriate basis for estimating PAEE. In particular, using a PAQ in the context of longitudinal studies might have the actively negative consequence of exaggerating the extent of the regular PA necessary to achieve a given long-term health outcome. Furthermore, this exaggeration might reduce the degree of population compliance with public health guidelines for desirable levels of PA.

This in turn suggests the opportunity to conduct new, large scale, randomly-sampled, nationally-representative diary studies, in which diarists complete the standard IPAQ and also carry instruments which objectively register their PA in real time (we suggest using accelerometers supplemented by heart-rate monitors, and perhaps GPS). These would allow a limited scope investigation, focussed on moderate-to-vigorous physical activity (MVPA) of any residual doubts about bias related to the selectivity of the CAPTURE-24 sample. More importantly, they would provide a means for calibrating METs attributions to the daily activities of representative samples in real daily practice compared with purposive samples in laboratory settings. It would also be possible to engage a subset of diary respondents (perhaps completing 7-day versions of the diary) in doubly-labelled water tests, thus providing a complete chain of evidence connecting the time-diary records to direct measurement of PAEE.

There are problems with the sorts of time-use diaries discussed here: participant burden is higher with time-use diaries than with passive observation methods such as cameras and accelerometers; the 10 minute intervals used by the HETUS standard are too coarse to capture some activities (leading to confusions between multiple short activities and simultaneously occurring activities within the same interval); and a single 24-hour coverage cannot represent 'usual' behaviour at an individual level. PAQ approaches can be used alongside diaries, to adjust diary estimates for longer term participation frequencies, and in turn to calibrate PAQ results to compensate for their biases (Gershuny 2012). The message of this current study is that diaries produce reliable results and should be used either alongside, or instead of, PAQ methods.

ACKNOWLEDGEMENTS

We thank all participants for volunteering for this research and to acknowledge the support of the UK Economic and Social Research Council (grant number ES/L011662/1 t). We also thank Sven Hollowell who helped prepare accelerometer data for this manuscript. The analysis was supported by the British Heart Foundation Centre of Research Excellence at Oxford (http://www.cardioscience.ox.ac.uk/bhf-centre-of-research-excellence) [grant number RE/13/1/30181 to AD], and the Li Ka Shing Foundation (http://www.lksf.org/) [to AD] and by an ERC Advanced Grant (Grant number 339703) [to JG]. We would also like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (http://dx.doi.org/10.5281/zenodo.22558). CF was funded by the British Heart Foundation Grant (BHF/PG/03/045).

References

Ainsworth, B. W Haskell, S Herrmann, N Meckes, D Bassett and C Tudor-Locke C. 2011. "Compendium of Physical Activities: A second Update of Codes and MET Values." Medicine Sciencs Sports Exercise 43:1575–1581.

Barker, Roger. 1968. "Ecological Psychology: Concepts and Methods for Studying the Environment of Human Behavior". Stanford, CA: Stanford University Press.

Barker, Roger and Associates. 1978. "Habitats, Environments, and Human Behavior". San Francisco, CA. Jossey-Bass.

Barker, Roger (ed.). 1963. "The Stream of Behavior: Explorations of its Structure and Content." New York: Appleton-Century-Crofts.

Bechtel, R., C Achepohl and R Akers. 1972. "Correlation Between Observed Behavior and Questionnaire Responses in Television Viewing" pp. 274-344 in Television and Social Behavior: Television in Day to Day Life: Patterns of Use, edited by EL Rubinstein, GA Comstock and JP Murray. Washington, DC: Government Printing Office.

Bernstein, Robert. A Chadha and R Montjoy. 2001. ''Over-Reporting Voting: Why It Happens and Why It Matters." Public Opinion Quarterly 65:22–44.

Brenner, Philip S. and John DeLameter. 2016. "Lies, Damned Lies, and Survey Self-Reports. Identity as a Cause of Measurement Bias." Social Psychology Quarterly 79: 333–354.

Brenner, Philip S., and John DeLamater. 2014. ''Social Desirability Bias in Self-Reports of Physical Activity: Is an Exercise Identity the Culprit?" Social Indicators Research 117:489–504.

Brunner, E., M Juneja and M Marmot. 2001. "Dietary Assessment in Whitehall II: Comparison of 7 D Diet Diary and Food-Frequency Questionnaire and Validity against Biomarkers." British Journal of Nutrition 86: 405–414.

Cowburn, G., A Matthews, A Doherty, A Hamilton, P Kelly, J Williams, C Foster and M Nelson. 2016. "Exploring the Opportunities for Food and Drink Purchasing and Consumption by Teenagers During their Journeys Between Home and School: A Feasibility Study Using a Novel Method. Public Health Nutrition. 19: 93–103.

Craig L. and Killian Mullan. 2011. "How Mothers and Fathers Share Childcare: A Cross-National Time-Use Comparison." American Sociological Review 76: 834–861.

Doherty A, SE Hodges, AC King, AF Smeaton, E Berry, CJA Moulin, S Lindley, P Kelly and C Foster. 2013. "Wearable Cameras in Health: The State of the Art and Future Possibilities." American Journal of Preventive Medicine 44: 320–323.

Doherty A, D Jackson, N Hammerla, T Plötz, P Olivier, M Granat, T White, V van Hees, M. Trenell, C Owen, S. Preece, R Gillions, S Sheard, T Peakman, S Brage and N Wareham. 2017. "Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study." PLoS ONE 12(2): e0169649.

Doherty, Aiden R., C Moulin, A Smeaton. 2011. “Automatically Assisting Human Memory: A SenseCam Browser.” Memory: Special Issue on SenseCam: The Future of Everyday Research? 19: 785–795.

Eurostat. 2009. Harmonised European Time Use Study Guidelines. Luxembourg: Eurostat.

Fisher, K., S. Chatzitheochari, E Gilbert, L Calderwood, E Fitzsimons, A Cleary, T Huskinson and J Gershuny. 2015. “A Mixed-Mode Approach to Measuring Young People’s Time Use in the Millennium Cohort Study.” Electronic International Journal of Time Use Research 12: 174–180.

Glorieux, I. and J Minnen. (2009). “How Many Days? A Comparison of the Quality of Time-Use Data from 2-Day and 7-Day Diaries.” Electronic International Journal of Time Use Research 6: 314–327.

Gershuny, J. and O Sullivan. 2017. United Kingdom Time Use Survey, 2014-2015. Centre for Time Use Research, University of Oxford. UK Data Service. SN: 8128, http://doi.org/10.5255/UKDA-SN-8128-1.

Gershuny, J. 2012. “Too Many Zeros: a Method for Estimating Long-term Time-use from Short Diaries.” Annals of Economics and Statistics 105-106: 247–271.

Goldschmidt-Clermont, L. and E Pagnossin-Aligisakis. 1999. “Households’ Non-SNA Production: Labour Time, Value of Labour and of Product, and Contribution to Extended Private Consumption.” Review of Income and Wealth Series 45: 519–529.

Harms, Teresa. 2004. “The Day at Home and Away: How Sixteen Danish Five-Year-Olds Spend their Time.” PhD Thesis, Department of Psychology, Roskilde University, Denmark.

Hill, M. 1985. “Patterns of Time Use”, pp 133-176 in Time, Goods and Well-Being, edited by F. Juster and F. Stafford. Ann Arbor, MI, Institute for Social Research, University of Michigan.

I-Min Lee, Eric., J Shiroma, Felipe Lobelo, Pekka Puska, Steven N Blair and Peter T Katzmarzyk. 2012. “Effect of Physical Inactivity on Major Non-Communicable Diseases Worldwide: An Analysis of Burden of Disease and Life Expectancy.” The Lancet: 380: 219–229.

Juster, Frank. 1985. “The Validity and Quality of Time Use Estimates Obtained from Retrospective Diaries”, pp 63-92 in Time, Goods and Well-Being, edited by F. Juster and F. Stafford. Ann Arbor, MI, Institute for Social Research, University of Michigan.

Man Yee Kan and Stephen Pudney. 2008. “Measurement Error in Stylized and Diary Data on Time Use.” Sociological Methodology. 38: 101–132.

Kelly P, A Doherty A, Mizdrak, S Marshall, J Kerr, A Legge, S Godbole, H Badland, M Oliver and C Foster. 2014. “High Group Level Validity but High Random Error of a Self-Report Travel Diary, as Assessed by Wearable Cameras.” Journal of Transport and Health 3: 190–201.

Kelly, P., S Marshall, H Badland, K Kerr, M Oliver, A Doherty and C Foster. 2013. “An Ethical Framework for Automated, Wearable Cameras in Health Behavior Research.” American Journal of Preventive Medicine 44: 314–319.

Kelly, P., E Thomas, A Doherty, T Harms, Ó Burke, J Gershuny and C Foster. 2015. "Developing a Method to Test the Validity of 24 Hour Time Use Diaries Using Wearable Cameras: A Feasibility Pilot." PLoS ONE 10: e0142198. https://doi.org/10.1371/journal.pone.0142198

Lee P., DJ Macfarlane, T Lam and SM Stewart. 2011. Validity of the international physical activity questionnaire short form (IPAQ-SF): A systematic review. The International Journal of Behavioral Nutrition and Physical Activity. 8:115.

Lesnard, L. 2010. "Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns." Sociological Methods and Research 38: 389–419.

O' Loughlin, G., S Cullen, A Goldrick, S O'Connor, R Blain, S O'Malley and G Warrington. 2013. "Using a Wearable Camera to Increase the Accuracy of Dietary Analysis." American Journal of Preventive Medicine 44: 297–301.

Pearson, Helen. 2015. "The Time Lab", Nature 526, 22.

Presser, S and L. Stinson. 1996. "Estimating the Bias in Survey Reports of Religious Attendance". Proceedings of the American Association for Public Opinion Research.

Robinson J. 1985. "The Validity and Reliability of Diaries Versus Alternative Time Use Measures", pp. 289–312 in Time, Goods and Well-Being, edited by F. Juster and F. Stafford. Ann Arbor, MI, Institute for Social Research, University of Michigan.

Robinson, John P. and P Converse. 1972. "Social Change Reflected in the Use of Time", pp. 17–86 in The Human Meaning of Social Change, edited by Angus Campbell and P Converse. New York, NY: Russell Sage Foundation.

Robinson, John. and Geoffrey Godbey. 1997. Time for Life: The Surprising Ways that Americans Use their Time. University Park, PA: Pennsylvania State University Press.

Shephard, R. J. 2003. ''Limits to the Measurement of Habitual Physical Activity by Questionnaires."British Journal of Sports Medicine 37:197–206.

Sullivan, O. 2000 "The Division of Domestic Labour: Twenty Years of Change?" Sociology 34: 437–456.

Szalai, A. (ed.) 1972. "The Use of Time: Daily Activities of Urban and Suburban Populations in Twelve Countries." The Hague: Mouton.

Tudor-Locke, C., T Washington, B Ainsworth and R Troiano. 2009. "Linking the American Time-Use Survey (ATUS) and the Compendium of Physical Activities: Methods and Rationale." Journal of Physical Activity and Health 6:347–353

Van Hees, V., S Sabia S, K Anderson, S Denton, J Oliver, M Catt. 2015. "A Novel, Open Access Method to Assess Sleep Duration Using a Wrist-Worn Accelerometer." PLoS ONE 10(11): e0142533. https://doi.org/10.1371/journal.pone.0142533.

Voorpostel M, T van der Lippe and J Gershuny 2009. "Trends in Free Time with a Partner: A Transformation of Intimacy?" Social Indicators Research 93:165–169.

[i] **C**omparing **A**nnotated **P**ictures with **T**ime-**U**se Diaries' **R**ecording of **E**vents over **24**-hours (CAPTURE-24).

[ii] "…the characteristic of a set of test scores that relates to the amount of random error from the measurement process that might be embedded in the scores". https://en.wikipedia.org/wiki/Reliability_(statistics) at 18/9/2017.

[iii] The 1-digit main categories are: (1) personal care; (2) employment; (3) household and family care; (4) voluntary work and meetings; (5) social life and entertainment; (6) sports and outdoor activities; (7) hobbies and computing; (8) mass media and; (9) travel and unspecified time use. A small number of additional codes were added to the Eurostat list to cope with ramera-related issues (eg "too dark to recognise").

[iv] The authors used this technique to compare total time-use patterns for pairs of countries.

[v] Including physical exercise to this analysis raises the correlation to .999.

[vi] Full tabulations of the regression results are in the Appendix.

[vii] "The **Metabolic Equivalent of Task (MET)**, or simply **metabolic equivalent**, is a physiological measure expressing the energy cost of physical activities and is defined as the ratio of metabolic rate (and therefore the rate of energy consumption) during a specific physical activity to a reference metabolic rate, set by convention to 3.5 ml $O_2 \cdot kg^{-1} \cdot min^{-1}$ …" https://en.wikipedia.org/wiki/Metabolic_equivalent at 18/9/2017