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; Millward and Spinney 2011; Spinney et al. 2011; van der Ploeg et al. 2010), 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, which tests self-report time-use diary reliability against objective criterion measures. The reliability of the camera and accelerometer evidence is unambiguous, as both instruments record aspects of respondents' activities in continuous real time. In this study, they are deployed as *criterion measures*—variables with self-evident reliability—as straightforward means of checking the duration of the activities recorded by respondents in their self-report time-use diaries.

Why not use the criterion variables rather than the diary measures? For some purposes (e.g. dietary analysis) wearable cameras are appropriate (O'Loughlin 2013), whilst for other topics (e.g. sleep) accelerometers are more suitable (van Hees et al. 2015). However, the camera records 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, whilst some activity categories (e.g. sleep, PA) can be inferred from accelerometer data, we are at present unable to identify other specific daily activities from PA evidence alone.

The underlying question is whether time-use diaries are an appropriate means of collecting data on durations of various types of activities. We start by deploying a large-scale survey (the 2014–15 UK National Time Use Study) (UK TUS) to compare estimates of participation rates in physical exercise from time-use diaries, with those derived from retrospective exercise participation questions from the same survey. The responses to retrospective participation questions are known to be seriously biased in directions determined both by respondents' perceptions of social desirability (Bauman et al. 2009; Bernstein Chadha and Montjoy 2001; Shepherd 2003; Troiano et al. 2012) 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 one widely used standard battery of such questions (IPAQ-SF) and reported that it seriously overestimated PA as measured by an objective criterion. Similarly, the 2014–15 UK TUS estimates show the retrospective questions produce participation rates approximately double those emerging from the diary records. Do time use diaries produce accurate estimates? Does the accuracy vary across different types of activities?

The immediate precursor to the current project was Kelly et al. 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 study extends this focus on travel behaviour, to include all daily activities—the entire range of paid and unpaid work, leisure, recreation, sleep and personal care activities. It is the first full-scale attempt to test the accuracy of continuous fully comprehensive diary records of adults against objectively registered, continuous and fully comprehensive measures of their daily activity recorded in real time.

2. Literature review

There is a long history of methodological research into time-use diary reliability studies—most examining the convergence of diaries with questionnaire-type time-use estimation methods, some comparing the diary with objective criterion variables.

The seminal work in the former category is the programme of work led by the Michigan Institute for Social Research, associated with the 1975 US National Time Use Study (Juster and Stafford 1985). 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), 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 speculation.

The CAPTURE-24 study follows the criterion variable route. The earliest direct test using a real-time activity record as an objective criterion deployed a video camera on top of a television set in 20 US households (Bechtel, Achepohl and Akers 1972). The camera record provided evidence of time in front of the television while switched on, for comparison with the diary record of television viewing. Anderson et al (1985) compared parents' reports of children's television viewing with a time lapse camera record of children's behaviour in front of the set. A second line of criterion comparison research uses motion sensors, worn continuously throughout the day, to compare with diary records of PA. Hofferth et al (2008) used this method to validate diary records of children's PA, as did van der Ploeg et al (2010) with a more general population-representative sample.

The present study uses both wearable cameras and accelerometers. It provides a substantial advance on the existing literature, yielding comprehensive comparisons of diary data with the criterion measures, covering all the activities of the day (rather than just television viewing or PA as in previous criterion-based studies). The diary/camera pairings directly compare durations in each daily activity, coded separately in the two records. The accelerometer data provide slightly less direct, but still comprehensive, comparisons, of the total daily PA estimated from the continuous accelerometer record, with estimates of the total daily PA in both the diary and the camera records achieved by attaching appropriate Metabolic Equivalent of Task scores (METs) to each diary/camera event (Tudor Locke et al 2009, Deyaert et al 2017). Although the focus of this paper is examining daily *durations*, this approach also provides some general testing (final paragraph Section 5.7 below) of the *timing* of activities during the day.

2.1 Estimating PA: Time-use data versus PAQs

Figure 1 (an updated version of Gershuny 2012: 258, which in turn follows discussion in Juster and Stafford 1985) uses the 2014–15 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 study, 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.

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 >0.5 probability of participation on a randomly chosen day (re-weighted, as in the previous paragraph). This type of reasoning gives us the 'predicted participation' line. Diary evidence on participation in walking, cycling, running and swimming provide participation rates 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 PA the previous month show some participation on the randomly chosen day. With these two exceptions, *all* of the diary participation rates are substantially below what would be expected from the questionnaire responses. 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 also supports findings from Lee 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 captured adequately by PAQ items. Respondents' detailed 'own words' diary descriptions provide continuous 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 understanding obesity, diabetes, cardiovascular disease and cancer (e.g. I-Min Lee et al. 2012) provide—in addition to the many social science applications mentioned above—a strong public health-based motivation for the time-use diary reliability evaluation enabled by the CAPTURE-24 project.

3. Study design and methods

3.1 Ethical considerations

The investigators developed a comprehensive ethical framework for conducting research using wearable cameras based on Kelly et al. (2013), and approved by the appropriate Oxford ethics committee (IDREC)[ii]. 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 or conversations. Participants were not permitted 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 sample varying broadly 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

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 study used the diary designed for the 2014–15 UK TUS, 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. Respondents were encouraged to record throughout the diary day, although, as in the majority of time-use surveys, most recording happened at the end of the day or early the following day. Typically, a one-day diary required 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 was developed by the Oxford Metrics Group (OMG), and 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 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. The Autographer is not waterproof, so participants were asked not to wear the camera if they were engaged in contact or water-based sports. 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 time-stamps 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.

Participants wore the accelerometer 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 most types of sport.

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 achieves 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) 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.

4 Data coding

The reliability test focus makes it essential to code the diary and image data independently. Limited resources allowed only a single coder for the own-words-descriptions of activities in the diary, so to avoid contamination, the diary and image coding exercises 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[iii]. The 10-minute interval makes it difficult for diarists to record brief (e.g. visiting the toilet, checking text messages) or momentary (e.g. taking medication, using an ATM) activities occupying less than 5 minutes. Episodes shorter than this sometimes fail to appear (although in some cases they 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 a 3-digit level[iv]. Primary and (up to three simultaneous) secondary activities are coded using the UK version of the standard HETUS activity classification (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 diaries coding procedures to the raw camera images, with two exceptions. First, we used a one-minute recording intervals, giving the image data a finer granularity than the diary. Second, the enjoyment domain was not used. For the comparisons discussed in the following sections, the one-minute intervals in the image files were concatenated to 10-minute diary intervals.

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 3: The SOP for image coding

We developed a standard operating procedure (SOP, Figure 3) 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’ (interruptions) 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 television → 2 frames of food preparation → 25 frames of watching television would be coded as a single activity *watching* television. If the food preparation lasted 3+ images, it would be coded as *preparing food* with *watching* television on either side (Figure 5 example). One of the limitations of the protocol is that it cannot assign either *preparing food* or *watching* television as primary or secondary activities unless it was recorded thus 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 (http://www.ukbiobank.ac.uk, Doherty et al. 2017). 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.

Accelerometer measures that represent total activity volume, such as average vector magnitude (i.e. movement per time interval relative to the centre of the earth), have been recommended as appropriate measures of PAEE. So to describe PA intensity, we aggregated the sample level data into ten-minute episodes for summary data analysis, maintaining the average vector magnitude value over the period (in milli-gravity units).

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 and significant differences in only three activity categories out of the total of 33 activities: *eating*, *reading* and *watching television*.

Table 2: Mean daily time in 33 activities

Figure 4 plots the 31 activity categories with durations less than 100 minutes. We have excluded *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 take just 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^{o} plot, with the divergence between the diary and questionnaire predictions in Figure 1.

Figure 4: 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 seem to be self-cancelling (i.e. with expected value zero across the sample), so as to produce the unbiased mean estimates seen in Figure 4.

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 2 show significant differences only in the case of time devoted to *eating*, *other personal care*, *food management*, *reading* and *school travel*.

Table 2 also provides measures of the covariance (correlation coefficients) of the two measures. 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. Some short duration categories, *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), have correlations <.5. 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, sequences of activities that can be nested hierarchically (Barker 1963, Barker, 1968, Barker 1978, Harms 2004). From the diarist's perspective, other simultaneous activities (e.g. drinking tea, glancing at the newspaper) may occur *within*, and are 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 evident which activities were primary or secondary/simultaneous. In these cases, we made 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. Interpreting images from the wearer's perspective (i.e. facing outwards) leads to other problems. A respondent eating a meal may turn to talk 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 'talking' as a secondary activity.

Table 3: Time-reporting hierarchy as seen in the camera record (mins/day)

We illustrate these problems by considering the full accounts of three activities in the entire camera record (Table 3). *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 activity, 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 variations in the amounts of activity reported by respondents with different levels of educational attainment; 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 highly-educated respondents are more likely to under- or over-report particular sorts of activities in their diaries compared with the camera evidence. Table 4 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 that occupy a relatively large proportion 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 are missing.

Table 4: Is there a reporting bias from educational level?

Most of the larger activities in Table 4 show reasonable correspondence between the recording patterns of the higher- and lesser educated participants, differences reflecting mainly the expected education-related variation. Among these activities, *sleep*, *eating*, *paid work*, *cooking*, *reading* and *watching television*, show similar patterns of difference between the two 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 respondents. *Radio listening*, *resting*, *exercise* and *exercise-related travel* show higher levels of difference 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. The focus of this paper is on evaluation of aggregate durations in activities, and with the exception of a brief discussion in the following section, we reserve analysis of the similarity of *timings* of daily activities for discussion elsewhere. Instead we now consider the *overall daily totals* of time in activities, using the compositional distance measure proposed by Robinson and Converse (1972)[v], 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.

The ratio of the mean of the 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 5 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 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 and 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 (e.g. in Table 4).

Figure 5: 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 necessary daily activities, such as eating and 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 'lower quality' diaries (those with more than 60 minutes missing/unallocated time during the diary day, and with fewer than 25 diary episodes) leaves 100 'higher 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).

5.6 Aggregate comparisons of diary, camera and accelerometer measures

Table 5 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 'higher 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, categorising all with fewer than 70 minutes missing as ‘better’ diaries. The lower two panels provide equivalent measures comparing the diary with the accelerometer records.

Table 5: Comparison of accelerometer means for summary activities, by camera and diary

Two findings emerge with some clarity from Table 5. The first is that both the camera and diary records show the expected differences in PA between 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 ‘higher 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 6: Accelerometer means by 2-digit activity categories, ordered by camera scores

Table 6 compares the mean accelerometer scores, broken down by both the camera and 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 push the estimate upwards.[vi])

5.7 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 coding for each 10 minute time-slot through the 1440 minutes of the day, on the mean accelerometer score for that time-slot. The time-slot 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 indicators (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[vii]. Considering that much of the variation in PA relates to physiological, demographic and socio-economic variables (BMI, level of 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 time-slot.

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 the camera measure. We modelled this by allocating MET[viii] scores—using the Ainsworth Compendium (Ainsworth et al. 2011) as a reference—to the 3-digit HETUS activity classification (Eurostat 2009). Our process broadly duplicated the work carried out by Tudor-Locke et al. (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 7 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 paper.

Table 7: Diary and camera-based METs as predictors of accelerometer scores

Although the main objective of the study is to validate diary estimates of activity durations, this last result, combined with the similarity of accelerometer scores between camera and diary records seen in Tables 5 and 6, allows a direct chain of inference to establish the general accuracy of time-of-day of activities in the diary. The camera timings of activities are objectively recorded. Therefore, the “same components” finding implies also that the diary is identifying close to the same times of day for the activities as is the camera record.

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 the social scientific or the public health literature, that such a test, covering all the activities of daily life, has been carried out.

We demonstrate that self-report time-use diaries provide a reliable basis for the accurate estimation of time-use patterns, without evidence of 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 large 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 different 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 self-report time-use diaries against reasonably objective criterion measures *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 PAQ battery is an insufficient and perhaps inappropriate basis for estimating PAEE. In particular, using PAQ data within longitudinal studies might have the consequence of exaggerating the extent of regular PA necessary to achieve specific long-term health outcomes. This over-estimation might itself reduce population compliance with public health guidelines.

The sample studied here is in no sense representative of any specific population. Despite our efforts to recruit a broad base 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 no evidence of a systematic bias from this source.

There are issues with the type of time-use diary used in this study. Participant burden is higher with time-use diaries than with passive data collection devices such as cameras and accelerometers. Furthermore, the 10 minute intervals used by the HETUS standard time-use diary are too coarse to capture some activities, leading to ambiguity (e.g. multiple short activities versus simultaneously-occurring activities within the same time-slot). We acknowledge that a single 24-hour period cannot represent ‘usual’ behaviour at an individual level. However, PAQ approaches could 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 study is that time-use diaries produce reliable results and should be used either alongside, or instead of, PAQ methods.

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.

Bauman, Adrian, Barbara E. Ainsworth, Fiona Bull, Cora L. Craig, Maria Hagströmer, James F. Sallis, Michael Pratt, and Michael Sjöström. 2009. “Progress and pitfalls in the use of the International Physical Activity Questionnaire (IPAQ) for adult physical activity surveillance.” Journal of Physical Activity and Health 6: S5–S8.

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.

Deyaert, J , T. Harms, D. Weenas, J. Gershuny and I. Glorieux 2017 “Attaching metabolic expenditures to standard occupational classification systems: perspectives from time-use research BMC”, BMC Public Health, DOI 10.1186/s12889-017-4546-7

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, Kimberly, 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, Jonathan 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, Jonathan. 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 Paul, 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, Paul., 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, Paul, 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.

Millward, Hugh, and Jamie Spinney. 2011. “‘Active Living’ Related to the Rural-Urban Continuum: A Time-Use Perspective.” The Journal of Rural Health 27: 141–150.

Pedišić, Ž., D. Dumuid and T. Olds. 2017. “Integrating sleep, sedentary behaviour, and physical activity research in the emerging field of time-use epidemiology: definitions, concepts, statistical methods, theoretical framework, and future directions.” Kinesiology: International Journal of Fundamental and Applied Kinesiology, 49: 10–11.

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.

Spinney, Jamie EL, Hugh Millward, and Darren M. Scott. 2011. “Measuring active living in Canada: A time-use perspective.” Social Science Research 40: 685–694.

Sullivan, Oriel. 2000. “The Division of Domestic Labour: Twenty Years of Change?” Sociology 34: 437–456.

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

Troiano, Richard, P. K Kelley, Gabriel Pettee, Gregory J. Welk, Neville Owen, and Barbara Sternfeld. 2012. “Reported physical activity and sedentary behavior: why do you ask?” Journal of Physical Activity and Health 9: S68–S75.

Tudor-Locke, Catrine., 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 der Ploeg, Hidde, P. Dafna Merom, Josephine Y. Chau, Michael Bittman, Stewart G. Trost, and Adrian E. Bauman. 2010. “Advances in population surveillance for physical activity and sedentary behavior: reliability and validity of time use surveys.” American Journal of Epidemiology 172: 1199–1206.

Van Hees, V., S Sabia, K Anderson, S Denton, J Oliver and 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] IDREC (University of Oxford Inter-Divisional Research Ethics Committee) reference number: SSD/CUREC1A/13-262.

[iii] The domains are primary activity, secondary activity, co-presence, location or travel mode, technology use, and enjoyment.

[iv] 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 camera-related issues (e.g. ‘camera off’).

[v] The authors used this technique to compare total time-use patterns for pairs of countries. This measure is more appropriate for constant-total time-use data than the somewhat similar compositional measures recommended for the purpose by Pedišić et al. 2017.

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

[vii] Full tabulations of the regression results are available upon request.

[viii] One MET (Metabolic Equivalent of Task) is defined as 1 kcal/kg/hour and is roughly equivalent to the energy cost of sitting quietly (Ainsworth et al. 2011). Intensity categories are broadly defined as light (<3 METs), moderate (3–6 METs) and vigorous (>6 METs); light-intensity categories can be interpreted as sleeping activities (<1 MET) or sedentary/lying/sitting activities (≥1 and <3 METs) (Ainsworth et al. 2011).