

# 1 **1. Introduction**

## 2 **1.1 Background**

3 Time-use diary methods are used for a range of research purposes in the social sciences.  
4 Economists use diary data to estimate extended National Product measures, including the  
5 value of unpaid work (Goldschmidt-Clermont and Pagnossin-Aligisakis 1999). Sociologists  
6 employ them to investigate parenting practices (Craig and Mullan 2011), sociability  
7 (Voorpostel, van der Lippe and Gershuny 2009) and the division of domestic labour (Sullivan  
8 2000). Whilst diaries are used as a data collection method by some public and population  
9 health researchers (e.g. Brunner, Juneja and Marmot 2001; Millward and Spinney 2011;  
10 Spinney et al. 2011; van der Ploeg et al. 2010), they are not routinely employed to estimate  
11 the extent and distribution of time devoted to physical activity (PA) across large populations.  
12 Rather, the convention has been to use various forms of physical activity questionnaires  
13 (PAQ) that include a battery of items asking respondents to recall the number of times they  
14 participated in specific activities over a specified period (last week/month). One of the most  
15 routinely used PAQs is the International Physical Activity Questionnaire (IPAQ), or its Short  
16 Form (IPAQ-SF).

## 17 **1.2 Objectives**

18 This paper reports the results of the CAPTURE-24<sup>i</sup> project, which tests self-report time-use  
19 diary reliability against objective criterion measures. The reliability of the camera and  
20 accelerometer evidence is unambiguous, as both instruments record aspects of respondents'  
21 activities in continuous real time. In this study, they are deployed as *criterion measures*—  
22 variables with self-evident reliability—as straightforward means of checking the duration of  
23 the activities recorded by respondents in their self-report time-use diaries.

24 Why not use the criterion variables rather than the diary measures? For some purposes (e.g.  
25 dietary analysis) wearable cameras are appropriate (O’Loughlin 2013), whilst for other topics  
26 (e.g. sleep) accelerometers are more suitable (van Hees et al. 2015). However, the camera  
27 records involve substantial extra costs (i.e. a similarly funded diary study alone might have  
28 achieved ten or more times the sample size discussed in this paper). Furthermore, whilst some  
29 activity categories (e.g. sleep, PA) can be inferred from accelerometer data, we are at present  
30 unable to identify other specific daily activities from PA evidence alone.

31 The underlying question is whether time-use diaries are an appropriate means of collecting  
32 data on durations of various types of activities. We start by deploying a large-scale survey  
33 (the 2014–15 UK National Time Use Study) (UK TUS) to compare estimates of participation  
34 rates in physical exercise from time-use diaries, with those derived from retrospective  
35 exercise participation questions from the same survey. The responses to retrospective  
36 participation questions are known to be seriously biased in directions determined both by  
37 respondents’ perceptions of social desirability (Bauman et al. 2009; Bernstein Chadha and  
38 Montjoy 2001; Shepherd 2003; Troiano et al. 2012) and by their attempts to enact particular  
39 sorts of normatively sanctioned identities (Brenner and DeLammater 2014). Lee, Macfarlane,  
40 Lam and Stewart (2011) carried out a systematic review of the validity of one widely used  
41 standard battery of such questions (IPAQ-SF) and reported that it seriously overestimated PA  
42 as measured by an objective criterion. Similarly, the 2014–15 UK TUS estimates show the  
43 retrospective questions produce participation rates approximately double those emerging  
44 from the diary records. Do time use diaries produce accurate estimates? Does the accuracy  
45 vary across different types of activities?

46 The immediate precursor to the current project was Kelly et al. 2014, which compared travel  
47 behaviour recorded by participants (n=69) wearing an automated SenseCam wearable camera  
48 with their registrations in a UK National Travel Survey-type trip log for the same day. The

49 CAPTURE-24 study extends this focus on travel behaviour, to include all daily activities—  
50 the entire range of paid and unpaid work, leisure, recreation, sleep and personal care  
51 activities. It is the first full-scale attempt to test the accuracy of continuous fully  
52 comprehensive diary records of adults against objectively registered, continuous and fully  
53 comprehensive measures of their daily activity recorded in real time.

## 54 **2. Literature review**

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

58 The seminal work in the former category is the programme of work led by the Michigan  
59 Institute for Social Research, associated with the 1975 US National Time Use Study (Juster  
60 and Stafford 1985). Robinson and Godbey (1997), having reviewed a number of previous  
61 examples of this type of methodological research (e.g. Robinson 1985, Juster 1985, Hill  
62 1985), concluded that additional controlled studies needed to be undertaken to extend and  
63 refine the estimates. Subsequent, methodologically sophisticated approaches to non-criterion-  
64 based tests (e.g. Kan and Pudney 2008) reiterate the view that diary approaches can be  
65 regarded as a ‘gold standard’. In their review, Brenner and DeLamater (2016) report no  
66 definitive progress in establishing validity or reliability on grounds other than *a priori*.  
67 Without an adequate criterion variable, deductive arguments are mere speculation.

68 The CAPTURE-24 study follows the criterion variable route. The earliest direct test using a  
69 real-time activity record as an objective criterion deployed a video camera on top of a  
70 television set in 20 US households (Bechtel, Achepohl and Akers 1972). The camera record  
71 provided evidence of time in front of the television while switched on, for comparison with  
72 the diary record of television viewing. Anderson et al (1985) compared parents’ reports of

73 children's television viewing with a time lapse camera record of children's behaviour in front  
74 of the set. A second line of criterion comparison research uses motion sensors, worn  
75 continuously throughout the day, to compare with diary records of PA. Hofferth et al (2008)  
76 used this method to validate diary records of children's PA, as did van der Ploeg et al (2010)  
77 with a more general population-representative sample.

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

## 90 **2.1 Estimating PA: Time-use data versus PAQs**

91 Figure 1 (an updated version of Gershuny 2012: 258, which in turn follows discussion in  
92 Juster and Stafford 1985) uses the 2014–15 UK TUS (Gershuny and Sullivan 2017) shows  
93 the relationship between the reported rates of PA participation from the questionnaire  
94 completed by respondents in the UK study, and the participation rates that emerge from their  
95 randomly selected diary days (weighted to give an equal representation of days of the  
96 week)—a convergent reliability test.

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

104 About 5% of those who report no walking and 2% of those who report no purposive PA the  
105 previous month show some participation on the randomly chosen day. With these two  
106 exceptions, *all* of the diary participation rates are substantially below what would be expected  
107 from the questionnaire responses. The average slope of the swimming, exercise, cycling,  
108 sport, walking and running lines is about half-way between the x-axis and the prediction line,  
109 which corresponds well with Brenner and DeLammeter’s (2014) ‘double the actual’  
110 estimation and also supports findings from Lee et al. (2011).

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

112 Another serious shortcoming is the constrained range of coverage of most PAQ batteries. All  
113 daily activities involve some level of physical activity energy expenditure (PAEE), but the  
114 PAQ items only cover a limited subset of pre-specified activities. Some respondents’ main  
115 source of PAEE may be outside the range covered by the PAQ. For example, incidental daily  
116 moderate-to-vigorous activities (e.g. caring for babies and toddlers, home renovation,  
117 gardening) are not captured adequately by PAQ items. Respondents’ detailed ‘own words’  
118 diary descriptions provide continuous coverage across all daily activities, resulting in a  
119 better-balanced estimation of the extent of different types of PA, although not their intensity.

120 These two issues with the PAQ approach, together with the centrality of PA measurement to  
121 understanding obesity, diabetes, cardiovascular disease and cancer (e.g. I-Min Lee et al.  
122 2012) provide—in addition to the many social science applications mentioned above—a  
123 strong public health-based motivation for the time-use diary reliability evaluation enabled by  
124 the CAPTURE-24 project.

### 125 **3. Study design and methods**

#### 126 **3.1 Ethical considerations**

127 The investigators developed a comprehensive ethical framework for conducting research  
128 using wearable cameras based on Kelly et al. (2013), and approved by the appropriate Oxford  
129 ethics committee (IDREC)<sup>ii</sup>. Participants signed a consent form after a member of the  
130 research team had fully explained the study requirements. Investigators recommended that  
131 participants check in advance that friends, family, and co-workers understood the nature of  
132 the study and were happy for them to take part, and were also advised of places where  
133 wearing the camera may not be appropriate (e.g. changing rooms, banks and schools). All of  
134 the cameras were encrypted and did not record sound or conversations. Participants were not  
135 permitted to keep any copies of the images.

#### 136 **3.2 Sample and setting**

137 The volunteer sample was drawn from the UK county of Oxfordshire. The research team  
138 invited participants via professional networks, free online advertisements, posters, social and  
139 sport clubs, word of mouth from other participants, and emails to an authorised list of willing  
140 research volunteers provided by a market research agency. Every effort was made to recruit a  
141 sample varying broadly across sex, age (18 years and over) and educational level (Table 1).

142 The original sample of 148 participants returned 124 complete diary, camera and  
143 accelerometer records, and 131 diary/camera pairs.

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

### 145 **3.3 Design**

146 The study design and associated protocols were refined based on the pilot study findings  
147 (n=14) (Kelly et al. 2015). Participants met with a member of the research team before and  
148 after the data collection day. The purpose of the initial meeting was to explain the project  
149 purpose, gain written informed consent, complete a short demographic questionnaire  
150 (including self-reported height and weight to calculate body mass index (BMI)) and receive  
151 the three instruments (diary, camera and accelerometer) and instructions on how to use them.  
152 On the data collection day, participants completed a self-report time-use diary and wore the  
153 two passive data collection devices (camera and accelerometer). Shortly after the data  
154 collection day, participants met with a researcher for a post-data collection ‘reconstruction  
155 interview’ and to report their experience of wearing the devices and completing the time-use  
156 diary. Participants received a £20 High Street voucher after completing the interview.

### 157 **3.4 Instruments, devices and interview**

#### 158 **3.4.1 Time-Use Diary**

159 The study used the diary designed for the 2014–15 UK TUS, the UK version of the European  
160 Harmonised European Time Use Study (HETUS) (Eurostat 2009). The diary starts at 4:00 am  
161 and covers 24-hours, in 10-minute intervals, with three hours on each page (Figure 2).  
162 Participants completed the diary in their own words across six fields or ‘domains’: primary  
163 activity, secondary activity, co-presence, location or travel mode, technology use, and  
164 enjoyment. Respondents were encouraged to record throughout the diary day, although, as in  
165 the majority of time-use surveys, most recording happened at the end of the day or early the  
166 following day. Typically, a one-day diary required about 20 minutes to complete.

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

### 168 **3.4.2 Autographer Wearable Camera**

169 The Autographer wearable camera was developed by the Oxford Metrics Group (OMG), and  
170 evaluated in several papers (e.g. Doherty et al. 2013). Participants wore the Autographer (on  
171 a lanyard or clipped to their clothing) for as long as possible during their waking hours—  
172 generally after showering in the morning until preparing for bed in the evening. The camera  
173 captured images automatically at 20- to 30-second intervals (medium capture rate) from the  
174 wearer's point of view, but no sound was recorded. A privacy lens allowed participants to halt  
175 image recording temporarily.

176 On a typical day, the camera captures 1500-2,000 images and also records ambient  
177 temperature and light levels. The average 16-hour battery life is sufficient to cover waking  
178 hours for most participants. The Autographer is not waterproof, so participants were asked  
179 not to wear the camera if they were engaged in contact or water-based sports. The camera  
180 functions best in good lighting conditions (i.e. daytime and indoors with sufficient lighting).  
181 Travelling after dark (particularly in winter) can result in unclear or poor quality images.  
182 Occasionally, participants' clothing or hair can obscure the lens, or data may be lost when the  
183 camera is turned off for various reasons (e.g. for privacy or unintentionally).

### 184 **3.4.3 Axivity AX3 band accelerometer**

185 The AX3, first released in 2012, is a continuous logging accelerometer designed for a range  
186 of applications including PA monitoring and classification, motion analysis and medical  
187 research (Doherty et al. 2017). The AX3 is compliant with the OpenMovement data format,  
188 has sufficient memory for 14 days continuous logging at 100Hz (512MB), is waterproof to  
189 1.5 meters and includes temperature and light sensors. It has an in-built, accurate clock and  
190 calendar which time-stamps the recorded acceleration data

191 (axivity.com/files/resources/AX3). The AX3 has configurable sample rates, adjustable  
192 sensitivity and a low power mode. The sample rate of 400 Hz gives a battery life of 5 days.  
193 Participants wore the accelerometer for at least 24-hours on their dominant hand (wrist),  
194 although many wore it for a day before and after the diary day, which provided an additional  
195 two days of sleep data. As the AX3 has a long battery life and is robust and water-proof,  
196 participants were able to wear it while working, travelling, taking a bath or shower, sleeping  
197 and playing most types of sport.

#### 198 **3.4.4 'Reconstruction' interview**

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

#### 207 **4 Data coding**

208 The reliability test focus makes it essential to code the diary and image data independently.  
209 Limited resources allowed only a single coder for the own-words-descriptions of activities in  
210 the diary, so to avoid contamination, the diary and image coding exercises were carried out  
211 separately, approximately four months apart (first diaries, then images). The large number of  
212 respondents, combined with the anonymity of the records, meant that the coder had no means  
213 of connecting particular diaries with the corresponding image files.

#### 214 **4.1 Time-use diary coding**

215 The HETUS diary instrument uses 10-minute intervals ('time-slots'). A time-use *episode* is a  
216 sequence of time-slots through which there is no change in any of the six substantive  
217 domains<sup>iii</sup>. The 10-minute interval makes it difficult for diarists to record brief (e.g. visiting  
218 the toilet, checking text messages) or momentary (e.g. taking medication, using an ATM)  
219 activities occupying less than 5 minutes. Episodes shorter than this sometimes fail to appear  
220 (although in some cases they appear in the secondary activity field). The final coded diary  
221 data file comprises, for each study participant, a sequence of episodes of varying lengths,  
222 starting at 4am, with a total duration of 1440 minutes (Eurostat 2009).

223 The HETUS activity coding system is hierarchical to a 3-digit level<sup>iv</sup>. Primary and (up to  
224 three simultaneous) secondary activities are coded using the UK version of the standard  
225 HETUS activity classification (just under 300 different activities). Coders categorise the main  
226 and secondary activities, location/mode of transport and other domains, and determine the  
227 start and end time of these episodes.

#### 228 **4.2 Camera image coding**

229 We applied the diaries coding procedures to the raw camera images, with two exceptions.  
230 First, we used a one-minute recording intervals, giving the image data a finer granularity than  
231 the diary. Second, the enjoyment domain was not used. For the comparisons discussed in the  
232 following sections, the one-minute intervals in the image files were concatenated to 10-  
233 minute diary intervals.

234 The interview notes were essential to the coding process. Most participants had a few black  
235 or unclear images from using the privacy lens cover, inadvertently covering it with clothing  
236 or being in low-light conditions, so the interviewer needed to identify what the respondent  
237 was doing when this occurred. The main reasons for covering the lens or turning the camera

238 off were showering, reading confidential documents on the computer, attending medical  
239 appointments and collecting children from school. The interview notes also allowed the coder  
240 to include additional domain information such as secondary activities, location and the  
241 presence of others.

#### 242 *Figure 3: The SOP for image coding*

243 We developed a standard operating procedure (SOP, Figure 3) for the image coding to aid  
244 replicability. Activities were identified as episodes and assigned a HETUS code if they  
245 continued for 3+ images with no ‘breaks’ (interruptions) of more than 2 images. Activities  
246 that lasted fewer than 3 images were grouped with the activity immediately preceding them.  
247 For example: 10 images of watching television → 2 frames of food preparation → 25 frames  
248 of watching television would be coded as a single activity *watching* television. If the food  
249 preparation lasted 3+ images, it would be coded as *preparing food* with *watching* television  
250 on either side (Figure 5 example). One of the limitations of the protocol is that it cannot  
251 assign either *preparing food* or *watching* television as primary or secondary activities unless  
252 it was recorded thus in the interview notes.

#### 253 **4.3 Accelerometer data extraction**

254 For the accelerometer data processing, we followed procedures used by the UK Biobank  
255 accelerometer data processing expert group, including device calibration to local gravity, and  
256 resampling to 100Hz (<http://www.ukbiobank.ac.uk>, Doherty et al. 2017). We calculated the  
257 sample level Euclidean norm of the acceleration in x/y/z axes, and removed machine noise  
258 using a fourth order Butterworth low pass filter with a cut-off frequency of 20Hz. In order to  
259 extract the activity-related component of the acceleration signal, we removed one  
260 gravitational unit from the vector magnitude, with remaining negative values truncated to

261 zero. Device non-wear time was automatically identified as consecutive stationary episodes  
262 lasting for at least 60 minutes.

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

## 268 **5. Data analysis and results**

### 269 **5.1. Aggregate comparison of diary and camera records**

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

276 *Table 2: Mean daily time in 33 activities*

277 Figure 4 plots the 31 activity categories with durations less than 100 minutes. We have  
278 excluded *sleep* and *paid work*, both with long durations, as they would distort the view, as  
279 well as give a correlation coefficient indistinguishable from unity. It shows a very strong  
280 association between the two measures as estimators of time-use at the aggregate level. If we  
281 take just the 31 two-digit activities as cases, we arrive at a correlation coefficient, between  
282 the diary and camera estimates, of .975, which is a remarkably high level of association  
283 between a self-report estimate and a criterion measure. Compare, for example, this nearly 45°  
284 plot, with the divergence between the diary and questionnaire predictions in Figure 1.

285 *Figure 4: 31 activities <100 minutes*

## 286 **5.2 Individual-level comparisons of diary and camera reports**

287 The similarity between the aggregate means of this quite detailed activity list is not entirely  
288 surprising. For example, it may be generated by perfect recall of the *sequence* of yesterday's  
289 activities, combined with a random error term in the recall of the *start/finish* time of each  
290 element in the activity sequence. The errors seem to be self-cancelling (i.e. with expected  
291 value zero across the sample), so as to produce the unbiased mean estimates seen in Figure 4.

292 Next, we turn from the comparison of *aggregate* mean time in activities across the sample, to  
293 consider the patterns of difference between the diary and camera estimates of total time in the  
294 activity at an *individual* level (i.e. moving from between-individual to within-individual  
295 comparisons). The main issue, for the present purpose of assessing the reliability of the diary  
296 record, is whether we can find statistically significant differences between diary-based  
297 estimates of the individual's total time in various activity categories, and the estimates  
298 derived from the (criterion) camera record. The t-tests in Table 2 show significant differences  
299 only in the case of time devoted to *eating, other personal care, food management, reading*  
300 *and school travel*.

301 Table 2 also provides measures of the covariance (correlation coefficients) of the two  
302 measures. The correlation coefficients can provide an estimate of the extent of 'noise'  
303 associated with recall errors in the start/finish times of diary activities, although it is not clear  
304 what should be considered a 'good' correlation in this context. Some short duration  
305 categories, *other paid work-related* (mean 15 minutes in the camera record), *resting and time*  
306 *out* (mean 8 minutes) and *listening to radio and recordings* (2 minutes), have correlations  
307  $<.5$ . However, the major time-use categories ( $>60$  minutes per day in the diary record) *sleep*,

308 *paid work, social activity, watching television* all have correlations  $> .65$ . Of the 33 activity  
309 categories, nine have  $r \geq .9$ , seven  $\geq .8$ , and a further five have  $r \geq .7$ .

### 310 **5.3 Simultaneous activities and the construction of daily narratives**

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

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

322 All simultaneous activities reported in the diaries and interviews were coded. However, if the  
323 respondent did not nominate the primary activity in the reconstruction interview, it was not  
324 always evident which activities were primary or secondary/simultaneous. In these cases, we  
325 made judgements in order to reconstruct the respondent’s ‘behaviour stream’ in a logical  
326 sequence. However, our judgements may have differed from the diarist’s subjective  
327 understanding of the particular activity. Interpreting images from the wearer’s perspective  
328 (i.e. facing outwards) leads to other problems. A respondent eating a meal may turn to talk to  
329 her companion, causing the camera to face away from the plate for a few frames. The analyst,  
330 for lack of other evidence, may classify this as conversing, even though the respondent would  
331 classify the primary activity as ‘eating’, with ‘talking’ as a secondary activity.

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

333 We illustrate these problems by considering the full accounts of three activities in the entire  
334 camera record (Table 3). *Eating* as a primary activity occupies 55 minutes in the camera  
335 record compared with 74 minutes in the diary. If all the events in which eating is recorded as  
336 a secondary activity were reversed to place eating as the primary activity, then eating  
337 durations would double. Similarly, *watching television*, 75 minutes as a primary activity in  
338 the diary but only 64 in the camera, increases by 50% if television viewing events counted as  
339 secondary by the camera analyst are recoded as primary. *Reading*, by contrast, is frequently  
340 ancillary to other activities. For example, during a meal, a respondent may read the  
341 newspaper rather than converse. The newsprint may feature frequently in the images  
342 alongside the plate of food, but from the diarist's perspective, eating the meal is the main  
343 activity.

#### 344 **5.4 Are there reporting differences by educational levels?**

345 The issue here is not whether there are variations in the amounts of activity reported by  
346 respondents with different levels of educational attainment; plainly we expect such  
347 differences. Rather, the question we ask is whether there are substantial *differences in the*  
348 *differences* between the camera and diary. Put more directly, we need to establish whether  
349 highly-educated respondents are more likely to under- or over-report particular sorts of  
350 activities in their diaries compared with the camera evidence. Table 4 compares the ratios of  
351 camera minus diary differences as a percentage of the diary mean estimates of time in the  
352 activities. In this analysis, we emphasise activities that occupy a relatively large proportion of  
353 the average day. Activities occupying 30 or fewer minutes per day have a relatively large  
354 number of zero-scores, meaning that either the diary or the camera evidence are missing.

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

356 Most of the larger activities in Table 4 show reasonable correspondence between the  
357 recording patterns of the higher- and lesser educated participants, differences reflecting  
358 mainly the expected education-related variation. Among these activities, *sleep, eating, paid*  
359 *work, cooking, reading* and *watching television*, show similar patterns of difference between  
360 the two records. *Household upkeep, gardening* and *pet-care* show larger differences, although  
361 with the same sign on the errors. Only *shopping, social entertainment* and *leisure travel* show  
362 large discrepancies in different directions. Among the shorter-mean duration activities, *other*  
363 *paid-work-related, helping other households* and *playing games* show substantially lower  
364 estimates in the diary records relative to the camera estimates among the less well-educated  
365 respondents. *Radio listening, resting, exercise* and *exercise-related travel* show higher levels  
366 of difference among the less well-educated respondents.

### 367 **5.5 Self-similarity analysis of diary and camera records**

368 We now consider similarities in the *overall patterns of time-use* produced by the camera and  
369 diary pairs in a more holistic way. The focus of this paper is on evaluation of aggregate  
370 durations in activities, and with the exception of a brief discussion in the following section,  
371 we reserve analysis of the similarity of *timings* of daily activities for discussion elsewhere.  
372 Instead we now consider the *overall daily totals* of time in activities, using the compositional  
373 distance measure proposed by Robinson and Converse (1972)<sup>v</sup>, calculating Generalised  
374 Euclidean Distances (GEDs) between pairs of records. By considering each of the 33 activity  
375 categories as an independent dimension, we can define a 32-dimensional hypotenuse-  
376 equivalent, as the square root of the sum of the squared differences between the paired  
377 camera and diary estimates of total time in each activity. The resulting ‘self-similarity’  
378 measure is the GED between the two time-use measures for a single respondent.

379 We can also calculate a similar GED between each of the 131 diary records and the camera  
380 records of each of the *other* 130 respondents, producing ‘general similarity’ measures. The  
381 self- and general-similarity measures together provide a 131\*131 matrix of GEDs, each row  
382 corresponding to a diary record and each column to a camera record, with the major diagonal  
383 elements containing the self-similarity measures.

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

390 Figure 5 reorders the rows and columns of the matrix in ascending order of the 131 self-  
391 similarity scores and, for each case, plots the mean of the general similarity indicator, the  
392 self-similarity indicator, and the minimum GED for the appropriate row of the matrix. The  
393 GED scores for each subject, roughly speaking, represent the sum of the deviations between  
394 the 33 time-use totals from camera/diary pairs; a GED of 100 units represents an average 3-  
395 minute deviation for the 33 pairs, 200 represents a 6-minute average deviation, and so on.  
396 With the exception of the single worst case, the self-similarity distance is smaller than the  
397 mean of the general similarity scores. Likewise, the self-similarity distance for most of the  
398 first 100 or so re-ordered cases is also the minimum GED. Beyond this point we find an  
399 increasing number of cases where the overall time-use pattern in the diary record is more  
400 similar to someone else’s camera record than to the diarist’s own record.

401 As already noted, there are two likely explanations for the differences between the camera  
402 and diary pairs. The first is simply poor diary-keeping, which emphasises the importance of

403 checking diaries for missing data upon collection. The second is the difference between the  
404 respondent's own recorded sequence of primary activities and the more complex multiple-  
405 simultaneous-activity reality of the camera record, and the coder's decisions. Although  
406 beyond the scope of this paper, this can be tested by observing the effects of re-ordering the  
407 multiple simultaneous activities recorded by coders in the camera records (e.g. in Table 4).

408 *Figure 5: Comparison of similarity of diary/camera pairs and distance of diaries to means of*  
409 *all other camera records*

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

## 420 **5.6 Aggregate comparisons of diary, camera and accelerometer measures**

421 Table 5 groups the 33 two-digit activities into seven broad categories and compares the PA  
422 levels (accelerometer records in mg/minute) associated with each. The upper two panels of  
423 the table refer to the camera records. On the right are the means and standard deviations for  
424 all participants who completed diaries, and on the left the 'higher quality' diaries. Only a  
425 subset (n=124) of the camera and diary sample returned usable accelerometer data. In order  
426 to maintain adequate numbers, we used a slightly less stringent criterion for diary quality,

427 categorising all with fewer than 70 minutes missing as ‘better’ diaries. The lower two panels  
428 provide equivalent measures comparing the diary with the accelerometer records.

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

430 Two findings emerge with some clarity from Table 5. The first is that both the camera and  
431 diary records show the expected differences in PA between broad types of activity. For  
432 example, in all four quadrants of the table we find a roughly eightfold difference in PA  
433 between the *sleeping* and *exercise* categories. In particular, the same differentials emerge  
434 from the camera and diary records.

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

441 Although the precise mechanism is not clear, in both cases the less densely-recorded diary  
442 and camera sequences reveal somewhat more exercise. Perhaps, in these cases, activities such  
443 as running for a bus or taking the stairs (which might otherwise be classified in a leisure, paid  
444 work or travel category) were instead placed in one of the subcategories of exercises,  
445 therefore slightly reducing the ‘all participants’ mean PA in the former categories and  
446 substantially increasing it in the latter.

447 *Table 6: Accelerometer means by 2-digit activity categories, ordered by camera scores*

448 Table 6 compares the mean accelerometer scores, broken down by both the camera and diary  
449 classification of each activity for the more detailed 2-digit activity classification. The rows of  
450 the table are placed in ascending order of the diary-based accelerometer scores. The ordering

451 would differ only slightly—activities moving up or down by no more than a single rank—if it  
452 were reordered according to the equivalent camera coding. There is a correlation of .98  
453 between the scores derived from the camera- and diary-based coding. (We excluded scores  
454 for exercise from our calculation of this correlation because, as distinct outliers, they would  
455 push the estimate upwards.<sup>vi</sup>)

## 456 **5.7 Individual-level comparisons of diary, camera and accelerometer measures**

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

464 The simple OLS approach to this is a ‘dummy variable regression’, classifying each time-  
465 slot-case by a vector of 32 indicators (0/1) variables representing the activity categories, 31 of  
466 which are always set to zero. The 33<sup>rd</sup> ‘default’ activity category is represented by the case  
467 where none of the indicator variables are set to 1. The camera-based regression analysis of  
468 the whole dataset produces a multiple correlation (R) coefficient of 0.493, the diary only  
469 slightly less at 0.473<sup>vii</sup>. Considering that much of the variation in PA relates to physiological,  
470 demographic and socio-economic variables (BMI, level of fitness, age, sex, employment  
471 status, social class, etc.) that can vary almost-independently of the type of activity, these are  
472 reassuringly acceptable levels of association from the perspective of the reliability of the two  
473 alternative indicators (i.e. camera and diary) of the type of activity in the time-slot.

474 However, assessing the reliability of the diary using the camera record as a criterion indicator  
475 requires a slightly different approach. It is important to know whether the diary measure is  
476 explaining the *same part* of the variation of the accelerometer record as the camera measure.  
477 We modelled this by allocating MET<sup>viii</sup> scores—using the Ainsworth Compendium  
478 (Ainsworth et al. 2011) as a reference—to the 3-digit HETUS activity classification (Eurostat  
479 2009). Our process broadly duplicated the work carried out by Tudor-Locke et al. (2009) who  
480 applied this to the American Time Use Study (ATUS) activity lexicon. The raw correlations  
481 between the camera- and diary-derived METs scores on one hand, and the accelerometer  
482 measure on the other, are 0.518 and 0.500 respectively.

483 Table 7 provides multiple correlation scores for model 3, which deploys both camera and  
484 diary estimated METs to predict accelerometer scores. The relatively small increment of  
485 prediction gained by adding the camera METs above the diary METs suggests that both the  
486 camera and diary are explaining *the same components* of the variation in the accelerometer  
487 record. Adding descriptors of the respondents (e.g. age, sex and educational attainment)  
488 improves the model performance, but we reserve further modelling of METs for another  
489 paper.

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

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

497 **6. Discussion**

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

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

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

517 The final observations relate more specifically to methods for estimating PA in the context of  
518 public health research. Combining the generally supportive evaluation of the diary against the  
519 camera and accelerometer in the two criterion-variable-based assessments, with the poor  
520 convergent reliability exhibited in the camera/PAQ comparison illustrated in Figure 1, we

521 conclude that the PAQ battery is an insufficient and perhaps inappropriate basis for  
522 estimating PAEE. In particular, using PAQ data within longitudinal studies might have the  
523 consequence of exaggerating the extent of regular PA necessary to achieve specific long-term  
524 health outcomes. This over-estimation might itself reduce population compliance with public  
525 health guidelines.

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

530 There are issues with the type of time-use diary used in this study. Participant burden is  
531 higher with time-use diaries than with passive data collection devices such as cameras and  
532 accelerometers. Furthermore, the 10 minute intervals used by the HETUS standard time-use  
533 diary are too coarse to capture some activities, leading to ambiguity (e.g. multiple short  
534 activities versus simultaneously-occurring activities within the same time-slot). We  
535 acknowledge that a single 24-hour period cannot represent 'usual' behaviour at an individual  
536 level. However, PAQ approaches could be used alongside diaries, to adjust diary estimates  
537 for longer term participation frequencies, and in turn to calibrate PAQ results to compensate  
538 for their biases (Gershuny 2012). The message of this study is that time-use diaries produce  
539 reliable results and should be used either alongside, or instead of, PAQ methods.

540

541

542

543 **References**

- 544 Ainsworth, B. W Haskell, S Herrmann, N Meckes, D Bassett and C Tudor-Locke C. 2011.  
545 “Compendium of Physical Activities: A second Update of Codes and MET Values.”  
546 *Medicine Sciences Sports Exercise* 43:1575–1581.
- 547 Barker, Roger. 1968. “Ecological Psychology: Concepts and Methods for Studying the  
548 Environment of Human Behavior”. Stanford, CA: Stanford University Press.
- 549 Barker, Roger and Associates. 1978. “Habitats, Environments, and Human Behavior”. San  
550 Francisco, CA. Jossey-Bass.
- 551 Barker, Roger (ed.). 1963. “The Stream of Behavior: Explorations of its Structure and  
552 Content.” New York: Appleton-Century-Crofts.
- 553 Bauman, Adrian, Barbara E. Ainsworth, Fiona Bull, Cora L. Craig, Maria Hagströmer, James  
554 F. Sallis, Michael Pratt, and Michael Sjöström. 2009. “Progress and pitfalls in the use  
555 of the International Physical Activity Questionnaire (IPAQ) for adult physical activity  
556 surveillance.” *Journal of Physical Activity and Health* 6: S5–S8.
- 557 Bechtel, R., C Achepohl and R Akers. 1972. “Correlation Between Observed Behavior and  
558 Questionnaire Responses in Television Viewing” pp. 274–344 in *Television and  
559 Social Behavior: Television in Day to Day Life: Patterns of Use*, edited by EL  
560 Rubinstein, GA Comstock and JP Murray. Washington, DC: Government Printing  
561 Office.
- 562 Bernstein, Robert. A Chadha and R Montjoy. 2001. “Over-Reporting Voting: Why It  
563 Happens and Why It Matters.” *Public Opinion Quarterly* 65: 22–44.
- 564 Brenner, Philip S. and John DeLameter. 2016. “Lies, Damned Lies, and Survey Self-Reports.  
565 Identity as a Cause of Measurement Bias.” *Social Psychology Quarterly* 79: 333–354.

- 566 Brenner, Philip S., and John DeLamater. 2014. "Social Desirability Bias in Self-Reports of  
567 Physical Activity: Is an Exercise Identity the Culprit?" *Social Indicators Research*  
568 117: 489–504.
- 569 Brunner, E., M Juneja and M Marmot. 2001. "Dietary Assessment in Whitehall II:  
570 Comparison of 7 D Diet Diary and Food-Frequency Questionnaire and Validity  
571 against Biomarkers." *British Journal of Nutrition* 86: 405–414.
- 572 Cowburn, G., A Matthews, A Doherty, A Hamilton, P Kelly, J Williams, C Foster and M  
573 Nelson. 2016. "Exploring the Opportunities for Food and Drink Purchasing and  
574 Consumption by Teenagers During their Journeys Between Home and School: A  
575 Feasibility Study Using a Novel Method." *Public Health Nutrition* 19: 93–103.
- 576 Craig L. and Killian Mullan. 2011. "How Mothers and Fathers Share Childcare: A Cross-  
577 National Time-Use Comparison." *American Sociological Review* 76: 834–861.
- 578 Deyaert, J , T. Harms, D. Weenas, J. Gershuny and I. Glorieux 2017 "Attaching metabolic  
579 expenditures to standard occupational classification systems: perspectives from time-  
580 use research BMC", *BMC Public Health*, DOI 10.1186/s12889-017-4546-7
- 581 Doherty A, SE Hodges, AC King, AF Smeaton, E Berry, CJA Moulin, S Lindley, P Kelly  
582 and C Foster. 2013. "Wearable Cameras in Health: The State of the Art and Future  
583 Possibilities." *American Journal of Preventive Medicine* 44: 320–323.
- 584 Doherty A, D Jackson, N Hammerla, T Plötz, P Olivier, M Granat, T White, V van Hees, M.  
585 Trenell, C Owen, S. Preece, R Gillions, S Sheard, T Peakman, S Brage and N  
586 Wareham. 2017. "Large Scale Population Assessment of Physical Activity Using  
587 Wrist Worn Accelerometers: The UK Biobank Study." *PLoS ONE* 12(2): e0169649.

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

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

592 Fisher, Kimberly, S. Chatzitheochari, E Gilbert, L Calderwood, E Fitzsimons, A Cleary, T  
593 Huskinson and J Gershuny. 2015. "A Mixed-Mode Approach to Measuring Young  
594 People's Time Use in the Millennium Cohort Study." *Electronic International Journal*  
595 *of Time Use Research* 12: 174–180.

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

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

602 Gershuny, Jonathan. 2012. "Too Many Zeros: a Method for Estimating Long-term Time-use  
603 from Short Diaries." *Annals of Economics and Statistics* 105-106: 247–271.

604 Goldschmidt-Clermont, L. and E Pagnossin-Aligisakis. 1999. "Households' Non-SNA  
605 Production: Labour Time, Value of Labour and of Product, and Contribution to  
606 Extended Private Consumption." *Review of Income and Wealth Series* 45: 519–529.

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

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

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

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

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

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

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

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

- 633 Lee P., DJ Macfarlane, T Lam and SM Stewart. 2011. "Validity of the international physical  
634 activity questionnaire short form (IPAQ-SF): A systematic review." *The International*  
635 *Journal of Behavioral Nutrition and Physical Activity* 8: 115.
- 636 Lesnard, L. 2010. "Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-  
637 Temporal Patterns." *Sociological Methods and Research* 38: 389–419.
- 638 O' Loughlin, G., S Cullen, A Goldrick, S O'Connor, R Blain, S O'Malley and G Warrington.  
639 2013. "Using a Wearable Camera to Increase the Accuracy of Dietary Analysis."  
640 *American Journal of Preventive Medicine* 44: 297–301.
- 641 Millward, Hugh, and Jamie Spinney. 2011. "'Active Living' Related to the Rural-Urban  
642 Continuum: A Time-Use Perspective." *The Journal of Rural Health* 27: 141–150.
- 643 Pedišić, Ž., D. Dumuid and T. Olds. 2017. "Integrating sleep, sedentary behaviour, and  
644 physical activity research in the emerging field of time-use epidemiology: definitions,  
645 concepts, statistical methods, theoretical framework, and future directions."  
646 *Kinesiology: International Journal of Fundamental and Applied Kinesiology*, 49: 10–  
647 11.
- 648 Robinson J. 1985. "The Validity and Reliability of Diaries Versus Alternative Time Use  
649 Measures", pp. 289–312 in *Time, Goods and Well-Being*, edited by F. Juster and F.  
650 Stafford. Ann Arbor, MI, Institute for Social Research, University of Michigan.
- 651 Robinson, John P. and P Converse. 1972. "Social Change Reflected in the Use of Time", pp.  
652 17–86 in *The Human Meaning of Social Change*, edited by Angus Campbell and P  
653 Converse. New York, NY: Russell Sage Foundation.
- 654 Robinson, John. and Geoffrey Godbey. 1997. *Time for Life: The Surprising Ways that*  
655 *Americans Use their Time*. University Park, PA: Pennsylvania State University Press.

- 656 Shephard, R. J. 2003. "Limits to the Measurement of Habitual Physical Activity by  
657 Questionnaires." *British Journal of Sports Medicine* 37:197–206.
- 658 Spinney, Jamie EL, Hugh Millward, and Darren M. Scott. 2011. "Measuring active living in  
659 Canada: A time-use perspective." *Social Science Research* 40: 685–694.
- 660 Sullivan, Oriel. 2000. "The Division of Domestic Labour: Twenty Years of Change?"  
661 *Sociology* 34: 437–456.
- 662 Szalai, Alexander. (ed.) 1972. *The Use of Time: Daily Activities of Urban and Suburban  
663 Populations in Twelve Countries*. The Hague: Mouton.
- 664 Troiano, Richard, P. K Kelley, Gabriel Pettee, Gregory J. Welk, Neville Owen, and Barbara  
665 Sternfeld. 2012. "Reported physical activity and sedentary behavior: why do you  
666 ask?" *Journal of Physical Activity and Health* 9: S68–S75.
- 667 Tudor-Locke, Catrine., T Washington, B Ainsworth and R Troiano. 2009. "Linking the  
668 American Time-Use Survey (ATUS) and the Compendium of Physical Activities:  
669 Methods and Rationale." *Journal of Physical Activity and Health* 6:347–353.
- 670 van der Ploeg, Hidde, P. Dafna Merom, Josephine Y. Chau, Michael Bittman, Stewart G.  
671 Trost, and Adrian E. Bauman. 2010. "Advances in population surveillance for  
672 physical activity and sedentary behavior: reliability and validity of time use surveys."  
673 *American Journal of Epidemiology* 172: 1199–1206.
- 674 Van Hees, V., S Sabia, K Anderson, S Denton, J Oliver and M Catt. 2015. "A Novel, Open  
675 Access Method to Assess Sleep Duration Using a Wrist-Worn Accelerometer." *PLoS  
676 ONE* 10(11): e0142533. <https://doi.org/10.1371/journal.pone.0142533>.
- 677 Voorpostel M, T van der Lippe and J Gershuny. 2009. "Trends in Free Time with a Partner:  
678 A Transformation of Intimacy?" *Social Indicators Research* 93:165–169.

---

<sup>i</sup> Comparing Annotated Pictures with Time-Use Diaries' Recording of Events over 24-hours (CAPTURE-24).

<sup>ii</sup> IDREC (University of Oxford Inter-Divisional Research Ethics Committee) reference number: SSD/CUREC1A/13-262.

<sup>iii</sup> The domains are primary activity, secondary activity, co-presence, location or travel mode, technology use, and enjoyment.

<sup>iv</sup> 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').

<sup>v</sup> 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.

<sup>vi</sup> Including physical exercise to this analysis raises the correlation to .999.

<sup>vii</sup> Full tabulations of the regression results are available upon request.

<sup>viii</sup> 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 ( $\geq 1$  and <3 METs) (Ainsworth et al. 2011).