Survey Research Methods (2024)
Vol. 18, No. 3, pp. 263-272
https://doi.org/10.18148/srm/8308
European Survey Research Association

Does Feedback from Physical Activity Measurement Devices Influence Physical Activity? Evidence From a Randomized Controlled Trial

Htay-Wah Saw[1,2] · Arie Kapteyn[2] · Jill E. Darling[2]
[1]University of Michigan-Ann Arbor
[2]Center for Economic and Social Research, University of Southern California

Wearable electronic devices are increasingly used to collect physical activity (PA) data. Most wearables provide PA feedback to users. The feedback has the potential for biasing measurements as users may increase their PA level in response to the feedback. On the other hand, the feedback can also be a desirable property as it can be employed to promote PA across settings. Studies examining the causal feedback effect independent of other factors have been limited. This study analyzed the causal effect of PA feedback provided by wearables. We implemented 4 field experiments over a period of 8 months. We recruited participants from a probability-based internet panel and asked them to wear continuously: (i) a PA device that does not provide feedback for 7 consecutive days, (ii) a PA device that provides feedback for another 7 consecutive days, (iii) both devices for another 1 or 2 consecutive days. After 6 PM each experimental day, participants completed a short online survey asking them about their experiences participating in the study. Of 120 eligible participants assessed, 81 provided valid and complete data and 39 were lost to follow-up. Participants with valid and complete data ($n=81$) accumulated 7% more PA on a given day when they wore a PA device providing feedback relative to when they wore a PA device that does not provide feedback (p-value<0.001). The feedback effect was robust to the inclusion of additional factors that might influence PA. Use of research-grade PA devices that provide no feedback is warranted for studies whose primary goals are to collect population PA data with minimal measurement errors, while wearables with feedback are most suited for PA intervention studies. When using PA devices that provide feedback, one needs to be aware of the bias that may result.

Keywords: wearables; physical activity; sensor data; measurement

1 Introduction

Wearable electronic devices (wearables) are increasingly used to collect physical activity (PA) data (e.g., All of Us Research Program Investigators, 2019; Doherty et al., 2017). These devices include both triaxial accelerometry such as GENEActiv and consumer smart watches such as Fitbit and Apple Watch. Unlike surveys that mainly rely on self-reports for collecting PA data, wearables provide ecologically-valid, highly granular, and longitudinal data because, with wearables, in-situ PA and sedentary behaviors can be measured repeatedly and frequently. Ecologically-valid measurement means data are collected in subjects' natural settings as they go about their normal lives aimed at enhancing generalizability and applicability of research findings to real-world scenarios and minimizing recall errors associated with the use of conventional survey approaches of data collection. Combining wearable data with survey data has allowed researchers to answer novel research questions that are not possible with the use of survey data alone (Kapteyn et al., 2018; Lathia et al., 2017). For instance, Kapteyn et al. (2018) conducted an accelerometry experiment in the Netherlands, England, and U.S., collecting both self-reports and accelerometry data, and found systematic differences between subjective and objective measurements across subgroups. In particular, the authors found that respondents over 65 described themselves equally active as younger respondents, though accelerometry data showed them to be significantly less ac-

The online version of this article (https://doi.org/10.18148/srm/8308) contains supplementary material.

Corresponding author: Htay-Wah Saw, University of Michigan-Ann Arbor, 426 Thompson Street, Room 4050, Ann Arbor, MI, 48104-1248, USA (Email: htaywah@umich.edu)

tive. With wearables, data are collected passively, reducing the need to administer survey questions and thus decreasing respondent survey burden. Additionally, the use of wearables to collect objective PA data eliminates differential item functioning associated with survey questionnaires. Questionnaire-based PA data collected in population representative studies consistently show a discrepancy between self-reports and objectively-measured PA data in relation to respondent characteristics (Kapteyn et al., 2018).

Commercially available wearables come with various sizes, shapes, technological features, and various degrees of user engagement, ranging from research-grade accelerometers, which do not provide information to the wearer on the level of PA, to devices that come with numerous user-friendly and attractive features for fully engaging and communicating with users. Use of wearables in research studies is not limited to research-grade wearables (e.g., GENEActiv, ActiGraph); some of the consumer wearables mainly intended for commercial clients (e.g., Fitbit, Apple Watch) have also been used extensively in research studies. With that in mind, it is important for researchers to have a thorough understanding of the measurement implications of different devices before they can be deployed in research studies. All wearables, except the research grade accelerometers, provide PA feedback to users in various ways, which include: (i) displaying daily PA progress on the device monitors, (ii) alerting users when daily PA goals or recommended PA level are not met, (iii) sending congratulatory messages to users when daily PA milestones are achieved, and (iv) allowing users to share achievement of daily PA milestones with peer groups and social networks. The feedback features provided by PA devices has the potential for creating response effects in which users are more likely to increase their PA level in response to the feedback compared to a situation where they receive no feedback about their PA level. The aim of this study is to evaluate the potential measurement bias introduced by wearable feedback.

Most prior clinical and intervention studies combined wearables that provide feedback with behavioral and educational approaches such as counselling and behavioral support provided by professionally-trained specialists, financial incentives, educational materials, information sessions, text messaging, online and peer supports. In other words, feedback was a desirable attribute (Brickwood et al., 2019). These studies were mainly aimed at evaluating the efficacy of multi-faceted behavioral interventions in promoting PA. This makes it difficult to disentangle the contribution of wearable feedback alone in influencing PA (Brickwood et al., 2019; Cadmus-Bertram et al., 2015; Lyons et al., 2014; Lyons et al., 2017; Mercer et al., 2016; Thompson et al., 2014; Van der Walt et al., 2018). Further, drawing a firm conclusion across prior studies has been hampered by heterogeneous findings, limited generalizability of findings due to use of convenience samples, small sample sizes, differences in study designs, use of various devices, and short study durations (Coughlin & Stewart, 2016). Despite their widespread use in numerous prior studies, studies rigorously examining the response effects of wearables that provide PA feedback independent of other factors have been limited (Brickwood et al., 2019).

In this study, we report findings from a series of field experiments we implemented among representative samples of US adults aged 50 and above, where respondents both wore PA devices that provide PA feedback, and devices that do not.

2 Methods

2.1 Data and study population

The data used in the analysis are from the Understanding America Study (UAS) Physical Activity Feedback Experimental Study (2023). The feedback study data are publicly available, but data user registration is required before the data can be accessed. Participants for this study were drawn from the UAS (USC, n.d.). Established at the University of Southern California (USC) in 2014, the UAS is a probability-based internet panel of US-households of approximately 13,000 respondents aged 18 and above. Respondents answer surveys on a computer, tablet, or smart phone, wherever they are and whenever they wish to participate. Panel members are recruited through address-based sampling. Prior access to internet is not a pre-requisite to be in the panel; respondents without prior internet access are provided with a computer tablet and broadband internet subscription. Respondents answer surveys once or twice a month. Partly as a result of this, the UAS comprises a vast amount of background information on its respondents, including extensive measures of physical and mental health, income, labor force participation, cognitive functioning, and demographics. Our study participants were drawn from a subsample of 200 UAS respondents who were 50 years or older. Respondents needed to have a smart phone or a tablet and (i) agree to wear PA wearable devices for 16 consecutive days; (ii) agree to complete an online survey at the end of each study day; (iii) return the wearable devices upon completion of the study; (iv) provide informed consent. Each participant received $50 remuneration to compensate for his/her time participating in this study. Out of the 200 drawn from the UAS, 30 did not respond to the consent survey, 40 did not meet the conditions mentioned above and 10 declined to participate.

2.2 Experimental design

We conducted 4 field experiments over a period of 8 months, beginning in July 2019 and ending in February 2020. Subjects were randomly assigned to the four experiments. In each of the four experiments, we used two different devices: Fitbit Versa (Fitbit Inc., San Francisco, CA n.d.), a wrist-worn device that provides PA feedback, communicates with participants and allows them to self-monitor their daily PA progress; and GENEActiv (GENEActiv, UK n.d.), a research-grade tri-axial and wrist-worn accelerometer that provides no PA feedback. In all experiments, participants were instructed to wear: (i) Fitbit 24 h/day for 7 or 8 consecutive days; (ii) both Fitbit and GENEActiv for one or two days (respondents were instructed to wear both devices on the same arm); (iii) GENEActiv 24 h/day for 7 or 8 consecutive days for a total of 15 or 16 experimental days. In experiments 1 and 3, participants were instructed to wear Fitbit first and then GENEActiv, whereas in experiments 2 and 4, participated started wearing GENEActiv first and then Fitbit after that. In experiments 3 and 4, we increased the number of days respondents wore both devices from one to two days. Respondents were instructed to charge the Fitbit for about 30–45 min each day. GENEActiv has a 30-day battery life, thus required no charging. Table 1 presents the dates of the four experiments and the manner in which they differ. Fig. 1 shows the flowchart of study participants for the 4 experiments. Of 200 sampled respondents, 30 did not respond to the consent survey, 40 did not meet the inclusion criteria (i.e., age 50 or above and having a smart phone or a tablet) and 10 declined to participate. Of the remaining 120 eligible participants, 32 individuals were assigned to experiment 1, 25 to experiment 2, 31 to experiment 3, and 32 to experiment 4. Thirty-nine participants were lost to follow-up for various reasons. Eighty-one participants provided complete and valid data for analysis. Table A1 in the Appendix presents a comparison of demographic characteristics between participants and non-participants (i.e., those lost to follow-up). The last column in Table A1 shows there are no significant differences in demographic composition between participants and non-participants.

Table 1

Experimental design setups and dates for each of the four experiments

	Experiment							
	1		2		3		4	
Experimental day	Date (2019)	Device worn	Date (2019)	Device worn	Date (2019)	Device worn	Date (2020)	Device worn
1	Jul 17	F	Aug 26	G	Dec 4	F	Jan 29	G
2	Jul 18	F	Aug 27	G	Dec 5	F	Jan 30	G
3	Jul 19	F	Aug 28	G	Dec 6	F	Jan 31	G
4	Jul 20	F	Aug 29	G	Dec 7	F	Feb 1	G
5	Jul 21	F	Aug 30	G	Dec 8	F	Feb 2	G
6	Jul 22	F	Aug 31	G	Dec 9	F	Feb 3	G
7	Jul 23	F	Sep 1	G	Dec 10	F	Feb 4	G
8	*Jul 24*	*F + G*	*Sep 2*	*G + F*	*Dec 11*	*F + G*	*Feb 5*	*G + F*
9	Jul 25	G	Sep 3	F	*Dec 12*	*F + G*	*Feb 6*	*G + F*
10	Jul 26	G	Sep 4	F	Dec 13	G	Feb 7	F
11	Jul 27	G	Sep 5	F	Dec 14	G	Feb 8	F
12	Jul 28	G	Sep 6	F	Dec 15	G	Feb 9	F
13	Jul 29	G	Sep 7	F	Dec 16	G	Feb 10	F
14	Jul 30	G	Sep 8	F	Dec 17	G	Feb 11	F
15	Jul 31	G	Sep 9	F	Dec 18	G	Feb 12	F
16	–	–	–	–	Dec 19	G	Feb 13	F

Rows in italics represent experimental days when participants wore both devices concurrently
F Fitbit, *G* GENEActiv

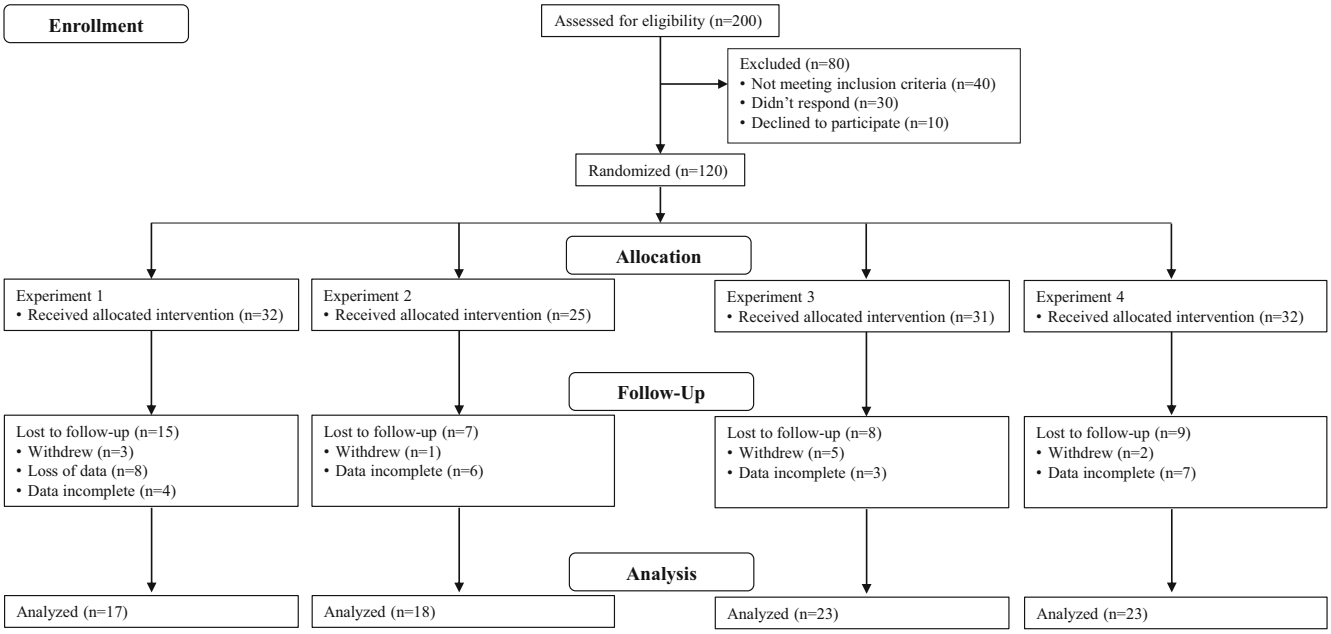

Fig. 1

Flowchart of study participants

2.3 End of day survey

After 6 PM each experimental day, participants were asked to log into their UAS accounts and complete a short "end of day" survey. The survey contained questions asking participants about: (i) the PA device they wore on a given day; (ii) if they took off the device; (iii) how many times they took off the device if they took it off; (iv) the times they took off and put the device back on; (v) the various reasons they took off the device; (vi) if they faced any issues and the kind of issues they faced wearing the devices; (vii) if they checked their PA level during the day using the Fitbit and how often they checked. Participants in experiments 1 & 2 and 3 & 4 completed a total of 15 and 16 end-of-day surveys, respectively.

2.4 PA measures

Our PA measures used in the analyses were in gravity (g) or acceleration units (1 g = 9.81 m/s^2) (Sabia et al., 2014). GENEActiv collects acceleration data at a frequency of 50 Hz, which we aggregated to one-minute totals. Fitbit provides PA data in step counts at one-minute intervals. Thus, if a respondent wore the devices continuously, both devices generated a total of 1440 data points on a given day. Both the GENEActiv and Fitbit provided non-wear data indicating whether participants wore the devices at a given one-minute interval or not. Using the GENEActiv and Fitbit data collected for the time periods when respondents wore both devices concurrently allowed us to calibrate the Fitbit steps with GENEActiv acceleration units, so that the data from both devices were in comparable gravity units. The method used for calibration consisted of running OLS regressions of GENEActiv acceleration units on Fitbit steps for every respondent separately. This allows for translating Fitbit steps into GENEActiv acceleration units by using the predicted GENEActiv PA corresponding to Fitbit steps. We then derived average PA measures for each respondent day. To do so, for each participant and for each device we summed up the PA data across each of the 7 days when respondents wore a device and then divided by total minutes the device was worn on that day. This yielded up to 7 average PA measures for each device and each participant, in addition to the days on which they were wearing both devices (used for calibration). We dropped data for any participant days for which the daily wear time was less than 1000 min (using different cut-offs for dropping the data did not alter the conclusions of our findings).

2.5 Statistical analysis

We first computed summary statistics for our study participants' demographics, end of day survey data, daily device wear time data, and PA data. Since our study participants were assigned to both treatment condition (i.e., wearing a Fitbit) and control condition (i.e., wearing a GENEAc-

Table 2

Summary statistics of study's participants demographic characteristics (%)

	Experiment				All experiments	*p*-value (test of demographic difference across experiments)
	1	2	3	4		
Gender						
Female	76(13)	67(12)	52(12)	48(11)	59(48)	0.239
Age						
Age (50–64)	53(9)	67(12)	57(13)	61(14)	59(48)	0.853
Age (65 and above)	47(8)	33(6)	43(10)	39(9)	41(33)	
Education						
GED or High School and below	12(2)	6(1)	17(4)	4(1)	10(8)	0.329
Some college	53(9)	44(8)	22(5)	35(8)	37(30)	
College and above	35(6)	50(9)	61(14)	61(14)	53(43)	
Number of participants	17	18	23	23	81	

Counts for each subgroup are in parentheses
GED (General Educational Development) is equivalent to a high school diploma in the United States
Individuals who did not complete a traditional high school program can obtain a GED by passing a series of tests that assess their knowledge and skills in core subject areas

tiv), while the order in which they wore the devices was randomly assigned, taking their individual differences in mean outcomes between Fitbit and GENEActiv wear periods and then taking the average of the mean differences across participants will generate an unbiased treatment effect. We first report our causal estimates of the feedback effect by taking the difference in mean outcomes between both conditions (Fitbit or GENEActiv) and then by performing a simple t-test of mean difference. We then run a series of multivariate OLS regressions to address the possibility that the feedback effect could have been driven by factors other than the PA feedback itself. The covariates that entered our OLS regressions included months in which the experiments were conducted (July, August, September, December, January, February), gender (male, female), age (50–64, 65+), education (GED, some college, college and above). Some participants might have worn the Fitbit during hours of the day when they were most active and conversely worn the GENEActiv during hours of the day when they were least active, which would spuriously generate higher PA for the Fitbit. To address any concerns that the feedback effect could also have been affected by selectivity of wear time, we report estimates for 5 OLS regression models restricting the sample to various wear times. Model 1 used the full sample. Models 2–5 restricted the sample to daily GENEActiv and Fibit wear time greater than: 1000 min/day, 1200 min/day, 1000 min/day for a minimum of 5 days, 1300 min/day, respectively. To account for clustering of our data (since we potentially have up to 14 data points for each participant in the regressions), we clustered the standard errors at the participant level. All analyses were conducted using Stata (version 17.0).

3 Results

Table 2 presents demographic characteristics of our study participants by experiment. The last column shows there are no significant differences in demographic composition across experiments. Overall, the ages of participants range from 50 to 83 years old with a mean age of 62. Fifty nine percent of participants are female, 10% have GED and below, 37% went to college but didn't earn any degrees, and 53% earned at least a college degree.

Table A2 in the Appendix presents summary statistics of participants' responses to end of day surveys for Fitbit and GENEActiv separately. For questions that are not applicable to the GENEActiv (e.g., if participants charged the device), "NAs" are inserted in the GENEActiv column. Device take-off rates vary by device type. Forty percent of participants reported taking off the Fitbit about once on a given day and the average take-off duration was 83 min. This compares with 11% of participants reporting taking off the GENEActiv once on a given day and the average take-off duration was 41 min. The main reason for taking off the Fitbit is "To charge the battery" (75%), and a small number of participants (9%) mentioned "To dry or clean the device" as a second reason. Eighty five percent of respondents chose "Other" as the main reason for taking off the GENEActiv. Although participants were instructed not to take off the GENEActiv when taking a shower, a majority of partici-

Table 3

Summary statistics of device wear time data: Average daily device wear time in minutes (hours)

Experiment	GENEActiv	Fitbit
1	1208 (20.13)	1223 (20.38)
2	1336 (22.27)	1092 (18.20)
3	1321 (22.02)	1222 (20.37)
4	1313 (21.88)	1130 (18.83)
Average	1323 (22.05)	1167 (19.45)

pants mentioned "To take a shower" as the main reason for taking off the GENEActiv under the "Other" response category (we did not include "To take a shower" as one of the possible response categories, as we were concerned that offering the response alternative might encourage the behavior). With respect to experience with wearing the devices, 94% and 88% of participants mentioned having no problems wearing the Fitbit and GENEActiv, respectively. Conditional on wearing the Fitbit, 65% of participants reported checking their PA level on a given day. Of those who reported checking, 44% mentioned they checked their PA once or twice a day, 25% four times a day, and 31% more than four times a day. Table 3 presents summary statistics for device wear time data. Across the four experiments, average daily wear times are 1323 min/day (22.05 h/day) and 1167 min/day (19.45 h/day) for GENEActiv and Fitbit, respectively.

Table 4 provides summary statistics of PA data and a quantitative estimate of the causal feedback effect. The first row shows summary statistics for the overall 14-days PA data (7 days when respondents wore GENEActiv only and 7 days when respondents wore Fitbit only); the overall PA mean value was calculated by summing up the PA data across 14 days, across respondents, and then divided by total minutes, considering only the time periods respondents wore the devices. The second and third rows provide similar PA data summary statistics for the 7-days period when participants wore the GENEActiv only and the 7-days period when participants wore the Fitbit only, respectively. The last row presents our estimate of the causal feedback effect, which is calculated as the mean PA difference between Fitbit and GENEActiv wear periods. The feedback effect is estimated at about 9.78 acceleration units (137.55–127.77), equivalent to 7% of overall PA average [(9.78/132.56) * 100], and statistically significant based on a t-test (p-value = 0.000). The PA difference between Fitbit and GENEActiv wear periods is equivalent to the treatment effect estimated from an unadjusted regression model.

Table 5 presents our multivariate OLS estimates with cluster robust variance estimates. We reported in Table 4 that the overall feedback effect was 9.78 acceleration units when we did not control for any potential cofounders (i.e., unadjusted model). As can be seen from Table 5 across the 5 models in the first row, the magnitude of the feedback effect remains largely unchanged, thus is robust to the inclusion of additional factors that might also affect the PA level. The robustness of the feedback effect estimate is mainly due to our longitudinal within-subject design, while the order in which participants wore the devices was randomized. This effectively eliminated the need to control for participants' demographic characteristics, as well as unmeasured and/or unmeasurable behavioral traits inherent across participants that could drive the PA level. Nevertheless, adding covariates to the model may increase precision of the feedback estimate, to the extent that the covariates reduce unexplained random variation across participants. Estimated coefficients for covariates have expected signs. PA level tends to be lower in Winter than in Fall; Females and the 65+ age group are less physically active than males and the 50–64 age group, respectively. Individuals with low education (GED or High School and below) are

Table 4

Summary statistics of PA data and quantitative estimate of feedback effect

	N (Participant days)	Mean	SD	Min	Max
(a) Overall 14-days period (7 days when respondents wore GENEActiv only and 7 days when respondents wore Fitbit only)	1065	132.56	45.11	22.02	407.40
(b) 7-days period when participants wore GENEActiv only	544	127.77	46.06	22.02	407.40
(c) 7-days period when participants wore Fitbit only	521	137.55	43.57	63.35	344.68

PA measurement was in gravity (g) or acceleration units (1 g = 9.81 m/s^2)
Feedback effect estimate = (Fibit – GENEActiv) = (c – b) = (137.55 – 127.77) = 9.78 acceleration units
t-test p-value = 0.000

Table 5

Average PA level multivariate regression analysis

	Average PA level									
	Model 1		Model 2		Model 3		Model 4		Model 5	
	Coef.	Std. Err	Coef.	Std. Err	Coef.	Std. Err	Coef.	Std. Err	Coef.	Std. Err
Device worn (Ref.: GENEActiv)										
Fitbit	11.953***	3.071	8.277***	2.813	8.131***	2.881	7.440**	3.007	8.977***	2.888
Month (Ref.: July)										
August	–13.012	13.685	–11.070	12.350	–8.376	11.837	–15.540	13.937	–8.288	11.278
September	–14.859	16.135	–6.791	15.192	–5.673	15.747	–6.646	16.608	–6.939	15.786
December	–44.649***	11.612	–38.832***	10.220	–37.499***	10.059	–39.491***	11.291	–37.668***	9.435
January	–19.026	12.758	–14.096	10.330	–17.696*	10.196	–10.515	11.631	–20.546**	9.682
February	–29.118**	11.963	–22.068**	10.253	–20.904**	10.062	–20.833*	11.488	–21.308**	9.611
Gender (Ref.: Male)										
Female	–15.008*	8.631	–8.950	7.224	–8.916	7.441	–8.133	8.077	–9.069	7.636
Age (Ref.: 50–64)										
Age 65+	–13.945*	7.717	–9.077	7.035	–8.799	7.143	–7.904	7.948	–10.067	7.271
Education (Ref.: GED)										
Some college	–8.721	15.619	–12.707	16.381	–18.999	15.318	–16.281	18.689	–17.519	14.937
College and above	–4.858	14.227	–2.204	15.543	–7.758	14.438	–2.783	17.633	–5.870	13.977
Constant	170.608***	17.859	160.278***	17.316	164.124***	16.453	161.577***	20.803	163.302***	15.924
Participant days (*N*)	1065		966		891		840		822	
R-squared	0.164		0.145		0.147		0.144		0.147	

Standard errors are cluster robust estimates to account for clustering of PA data at the participant level
Sample restrictions for each of the 5 models are the following
Model 1: Full sample
Model 2: Fitbit and GENEActiv daily wear time greater than 1000 min
Model 3: Fitbit and GENEActiv daily wear time greater than 1200 min
Model 4: Participants wore Fitbit and GENEActiv for at least 1000 min/day for at least 5 days
Model 5: Fitbit and GENEActiv daily wear time greater than 1300 min
PA measurement was in gravity (g) or acceleration units (1 g = 9.81 m/s2)
* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$

more active than individuals with some college education and above. Results are quantitatively as well as qualitatively similar if we instead estimate the regressions using linear hierarchical random-effects models (we present estimated coefficients from linear hierarchical random-effects models in Table A3 in the Appendix).

4 Discussion

This study implemented a series of field experiments over a period of 8 months among UAS panel members aged 50 and above to investigate if PA feedback from wearable devices leads users to increase their PA level. We find that on average participants accumulated 7% more PA in a given day when they wore the Fitbit Versa than when they wore the GENEActiv. Our estimate of the feedback effect is robust to whether we controlled for additional confounders that might drive PA level or not. Our results are consistent with a recent meta-analysis (Brickwood et al., 2019), which found a similar average effect size across clinical intervention studies mostly aimed at promoting PA among at-risk populations that combined wearables with traditional behavioral interventions such as education materials, group or individual counselling (either in-person or by telephone), information sessions, financial incentives, text messaging, dietary restrictions, and online support.

Our study generates results with measurement and methodological implications for future studies planning on using wearables for collecting PA data. Inactivity is one of the leading contributors to premature deaths, non-communicable diseases, disabilities, and accounts for about 10% of premature deaths in the U.S (Carlson et al., 2018; Carlson et al., 2015; Lee et al., 2012). A large percentage of the U.S. adult population remains physically inactive and do not meet PA guidelines recommended by the Centers for Disease Control and Prevention (CDC) (Blackwell & Clarke, 2018). For accurate diagnosis of population PA level and sedentary behavior, it is important that population PA data are collected using a measurement tool (or a combination of measurement tools) not affected by respondents' behavioral responses. Most large-scale studies of population PA have used self-report questionnaires for measuring population PA (Bauman et al., 2009; Farrell et al., 2014). However, prior studies reveal that compared to younger individuals, older adults tend to significantly over-report their PA level when questionnaires are used (Kapteyn et al., 2018). The use of wearables can help address measurement issues inherent in self-reports of PA. However, as our results show, use of wearables that provide feedback can generate response effects, resulting in an overestimate of actual PA level. Ideally, research-grade wearable devices that do not provide feedback would seem to be the best measurement approach if the primary goal is to accurately assess PA. However, in view of the increasing prevalence of commercial activity trackers in the general population, using these for measurement purposes is an attractive alternative. In that case, the response effects of the feedback provided by these devices need to be accounted for.

The main strengths of our study include: (i) recruitment of study participants from a random sample from the Understanding American Study (UAS)—a probability-based online panel representative of the U.S. adult population, and (ii) use of a longitudinal within-subject experimental design that enables us to generate robust and causal estimates of the feedback effect independent of other educational and behavioral interventions, of respondents' demographic and unmeasured and/or unmeasurable characteristics. Study limitations include the following. Our intervention period of 15–16 days was relatively short; thus, we are unable to ascertain whether the feedback effect found in our study will persist beyond the intervention period. Our study was not designed to disentangle various mechanisms that mediate or moderate the feedback effect, which would require a research design that is more elaborate than the current study.

Daily total PA as measured by wearables comes from three sources: leisure time PA, occupational PA, and commuting PA. Prior studies however show that not all PA is created equal. While the positive health benefits of leisure time PA across various health outcomes are well documented, evidence is still mixed for occupational PA with some studies indicating that occupational PA can have detrimental effects on mortality and some health outcomes (Abu-Omar & Rütten, 2008; Barengo et al., 2006; Beenackers et al., 2012; Blackwell & Clarke, 2018; Gutiérrez-Fisac et al., 2002; Oppert et al., 2006). Wearables by themselves are not able to distinguish between leisure time PA and occupational PA. Future studies should include survey questions asking participants about the time and duration they engage in various activities/tasks throughout the day. This will enable researchers to break down the overall feedback effect by various PA sources. Although the UAS is representative of the US adult population aged 18 and above, our study was limited to UAS respondents aged 50 and above having a smartphone. According to the Pew Research Center (Pew Research Center, 2021), about 83% of the US adult population owned a smartphone when the experiment was conducted in 2019–2020. Future research should also examine if the magnitude and direction of the feedback effects vary across study populations, socioeconomic subgroups, and lengths of the time period participants wear the device. Due to a relatively small sample size, our study was not able to meaningfully analyze potential mechanisms that moderate the feedback effect. Future studies should also

examine various demographic, behavioral, and contextual factors that moderate the feedback effect.

In conclusion, this study found that asking participants to wear a PA device that provides feedback led to an increase in PA level by 7%. The feedback effect is robust to the inclusion of additional factors that might influence PA. Wearable PA devices should be part of future large-scale population health and aging studies. However, the type of wearable devices to be used in research studies should be selected carefully taking into full consideration device characteristics, study goals and objectives.

Acknowledgements Informed consents were obtained from all participants. Protocols dictating conduct of this study were approved by the Institutional Review Board (IRB) at USC.
The study was funded by a grant from the National Institute on Aging and the Social Security Administration (grant number 5U01AG054580).

References

Abu-Omar, K., & Rütten, A. (2008). Relation of leisure time, occupational, domestic, and commuting physical activity to health indicators in Europe. *Preventive Medicine*, *47*(3), 319–323.

All of Us Research Program Investigators (2019). The "All of Us" research program. *New England Journal of Medicine*, *381*(7), 668–676.

Barengo, N.C., Kastarinen, M., Lakka, T., Nissinen, A., & Tuomilehto, J. (2006). Different forms of physical activity and cardiovascular risk factors among 24–64-year-old men and women in Finland. *European Journal of Cardiovascular Prevention & Rehabilitation*, *13*(1), 51–59.

Bauman, A., Bull, F., Chey, T., Craig, C.L., Ainsworth, B.E., Sallis, J.F., Bowles, H.R., Hagstromer, M., Sjostrom, M., & Pratt, M. (2009). The international prevalence study on physical activity: results from 20 countries. *International Journal of Behavioral Nutrition and Physical Activity*, *6*(1), 21.

Beenackers, M.A., Kamphuis, C.B., Giskes, K., Brug, J., Kunst, A.E., Burdorf, A., & Van Lenthe, F.J. (2012). Socioeconomic inequalities in occupational, leisure-time, and transport related physical activity among European adults: a systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, *9*(1), 116.

Blackwell, D.L., & Clarke, T.C. (2018). State variation in meeting the 2008 federal guidelines for both aerobic and muscle-strengthening activities through leisure-time physical activity among adults aged 18–64: United States, 2010–2015. *National Health Statistics Reports*, *112*, 1–22.

Brickwood, K.-J., Watson, G., O'Brien, J., & Williams, A.D. (2019). Consumer-based wearable activity trackers increase physical activity participation: systematic review and meta-analysis. *JMIR mHealth and uHealth*, *7*(4), e11819.

Cadmus-Bertram, L.A., Marcus, B.H., Patterson, R.E., Parker, B.A., & Morey, B.L. (2015). Randomized trial of a Fitbit-based physical activity intervention for women. *American Journal of Preventive Medicine*, *49*(3), 414–418.

Carlson, S.A., Fulton, J.E., Pratt, M., Yang, Z., & Adams, E.K. (2015). Inadequate physical activity and health care expenditures in the United States. *Progress in Cardiovascular Diseases*, *57*(4), 315–323.

Carlson, S.A., Adams, E.K., Yang, Z., & Fulton, J.E. (2018). *Peer reviewed: percentage of deaths associated with inadequate physical activity in the United States*. Preventing Chronic Disease, Vol. 15.

Coughlin, S.S., & Stewart, J. (2016). Use of consumer wearable devices to promote physical activity: a review of health intervention studies. *Journal of Environment and Health Sciences*. https://doi.org/10.15436/2378-6841.16.1123.

Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M.H., White, T., Van Hees, V.T., Trenell, M.I., & Owen, C.G. (2017). Large scale population assessment of physical activity using wrist worn accelerometers: the UK biobank study. *PloS ONE*, *12*(2), e169649.

Farrell, L., Hollingsworth, B., Propper, C., & Shields, M.A. (2014). The socioeconomic gradient in physical inactivity: evidence from one million adults in England. *Social Science & Medicine*, *123*, 55–63.

Fitbit Inc. (n.d.). Fitbit Versa 4 Smartwatch. https://www.fitbit.com/global/us/products/smartwatches/versa?sku=507BKBK. Accessed 29 Dec 2023.

GENEActiv (n.d.). GENEActiv: Raw data accelerometer. https://activinsights.com/technology/geneactiv/. Accessed 29 Dec 2023.

Gutiérrez-Fisac, J.L., Guallar-Castillón, P., Díez-Gañán, L., García, E.L., Banegas, J.R.B., & Artalejo, F.R. (2002). Work-related physical activity is not associated with body mass index and obesity. *Obesity Research*, *10*(4), 270–276.

Kapteyn, A., Banks, J., Hamer, M., Smith, J.P., Steptoe, A., Van Soest, A., Koster, A., & Wah, S.H. (2018). What they say and what they do: comparing physical activity across the USA, England and the Netherlands. *J Epidemiol Community Health*, *72*(6), 471–476.

Lathia, N., Sandstrom, G.M., Mascolo, C., & Rentfrow, P.J. (2017). Happier people live more active lives: Using smartphones to link happiness and physical activity. *PloS ONE*, *12*(1), e160589.

Lee, I.-M., Shiroma, E.J., Lobelo, F., Puska, P., Blair, S.N., Katzmarzyk, P.T., & Group, L. P. A. S. W. (2012). Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The Lancet*, *380*(9838), 219–229.

Lyons, E.J., Lewis, Z.H., Mayrsohn, B.G., & Rowland, J.L. (2014). Behavior change techniques implemented in electronic lifestyle activity monitors: a systematic content analysis. *Journal of Medical Internet Research*, *16*(8), e192.

Lyons, E.J., Swartz, M.C., Lewis, Z.H., Martinez, E., & Jennings, K. (2017). Feasibility and acceptability of a wearable technology physical activity intervention with telephone counseling for mid-aged and older adults: a randomized controlled pilot trial. *JMIR mHealth and uHealth*, *5*(3), e28.

Mercer, K., Li, M., Giangregorio, L., Burns, C., & Grindrod, K. (2016). Behavior change techniques present in wearable activity trackers: a critical analysis. *JMIR mHealth and uHealth*, *4*(2), e40.

Oppert, J., Thomas, F., Charles, M., Benetos, A., Basdevant, A., & Simon, C. (2006). Leisure-time and occupational physical activity in relation to cardiovascular risk factors and eating habits in French adults. *Public Health Nutrition*, *9*(6), 746–754.

Pew Research Center (2021). Mobile fact sheet. https://www.pewresearch.org/internet/fact-sheet/mobile/

Sabia, S., van Hees, V.T., Shipley, M.J., Trenell, M.I., Hagger-Johnson, G., Elbaz, A., Kivimaki, M., & Singh-Manoux, A. (2014). Association between questionnaire- and accelerometer-assessed physical activity: the role of sociodemographic factors. *American Journal of Epidemiology*, *179*(6), 781–790.

Thompson, W.G., Kuhle, C.L., Koepp, G.A., McCrady-Spitzer, S.K., & Levine, J.A. (2014). "Go4Life" exercise counseling, accelerometer feedback, and activity levels in older people. *Archives of Gerontology and Geriatrics*, *58*(3), 314–319.

UAS Physical Activity Feedback Experimental Study (2023). Produced by the USC Dornsife Center for Economic and Social Research, with funding from the National Institute on Aging and the Social Security Administration. https://uasdata.usc.edu/index.php?r=eNpLtDKyqi62MrFSKkhMT1WyLrYyNAeyS5NyMpP1UhJLEvUSU1Ly80ASQDWJKZkpUKahoaGpknUtXDB_uxMO. Accessed 29 Dec 2023.

USC (n.d.). Understanding America Study (UAS). https://uasdata.usc.edu/index.php. Accessed 29 Dec 2023.

Van der Walt, N., Salmon, L.J., Gooden, B., Lyons, M.C., O'Sullivan, M., Martina, K., Pinczewski, L.A., & Roe, J.P. (2018). Feedback from activity trackers improves daily step count after knee and hip arthroplasty: a randomized controlled trial. *The Journal of Arthroplasty*, *33*(11), 3422–3428.