Received: 10 June 2025 | Accepted: 9 February 2026

DOI: 10.1111/bjhp.70063

the british psychological society
promoting excellence in psychology

ARTICLE

Do adults with cardiovascular disease risk show meaningful reactivity to physical activity measurement? Coordinated analysis across six studies

Danielle Arigo[1,2] | **Kiri Baga**[1] | **Amanda L. Folk**[1] | **Gabrielle M. Salvatore**[1] | **Iris Bercovitz**[1] | **Ria Singh**[1] | **Laura M. König**[3] | **Meghan L. Butryn**[4] | **Jacqueline A. Mogle**[5]

[1]Department of Psychology, Rowan University, Glassboro, New Jersey, USA

[2]Department of Family Medicine, Rowan-Virtua School of Osteopathic Medicine, Stratford, New Jersey, USA

[3]Faculty of Psychology, University of Vienna, Vienna, Austria

[4]Department of Psychology, Drexel University, Philadelphia, Pennsylvania, USA

[5]RTI Health Solutions, Durham, North Carolina, USA

Correspondence
Danielle Arigo, Department of Psychology, Rowan University, 201 Mullica Hill Road, Robinson Hall 116G, Glassboro, NJ 08028, USA.
Email: arigo@rowan.edu

Funding information
National Heart, Lung, and Blood Institute

Abstract

Objectives: To estimate the extent of physical activity (PA) measurement reactivity among adults ages 40–60 with risk factors for cardiovascular disease (CVD), to inform best practices for addressing reactivity in PA research and intervention.

Design: Coordinated secondary analysis across six datasets from studies that used 6–7 days of observation following the introduction of PA measurement devices. Moderators of interest were demographic and study design characteristics.

Methods: We included data from participants ages 40–60 with ≥1 CVD factors who provided device-assessed PA behaviour across 6–7 days ($N = 1825$). We used multilevel modelling to examine participants' PA behaviour (i.e., activity units, steps per day) across days, with decreases in activity indicating reactivity. The threshold for statistical significance was set at $p < .05$ and standardized effect sizes of interest were semipartial correlation coefficients (*sr*s) ≥.25; we also report conversions to Cohen's *d* and corresponding equivalence tests.

Results: No patterns met both criteria for significance for either main or moderation effects, including tests of study design features. Results from one small study showed a decrease in steps per day across days of observation ($p = .15$,

sr = .26, *d* = .23, 90% CI: −.03, .50), though men showed an *increase* in steps per day (whereas women decreased).
Conclusions: Adults in midlife with CVD risk factors show little evidence of PA measurement reactivity. We recommend continuing to examine PA patterns in individual studies, though widespread use of burdensome procedures to prevent reactivity is not warranted in this at-risk population.

KEYWORDS
cardiovascular risk, gender difference, intensive assessment, measurement reactivity, midlife, physical activity, sex difference

Statement of Contribution

What is already known on this subject?

- Equivocal evidence regarding the effect of measurement reactivity on physical activity estimates.

What does this study add?

- Coordinated analysis found minimal measurement reactivity among adults with CVD risk factors.
- Burdensome methods such as adding observation days may be unwarranted.

INTRODUCTION

Measurement reactivity has gained considerable attention as a source of bias in physical activity (PA) measurement (Christiansen et al., 2023; Davis & Loprinzi, 2016; Hilden et al., 2023). Measurement reactivity is the change in behaviour due to its direct observation; with respect to PA, measurement increasingly involves monitoring devices such as Fitbit or pedometers (Davis & Loprinzi, 2016). Device-based PA measurement most often occurs over multiple days to capture observable fluctuations in activity, and daily PA totals are averaged to estimate 'typical' PA (Tudor-Locke et al., 2010). Reactivity is most often defined as greater PA early in the measurement period (vs. later), and some studies demonstrate that participants show more PA on initial days of observation than on subsequent days (Baumann et al., 2018; Clemes & Deans, 2012). This is thought to result from increased awareness of PA on initial days (due to the novel presence of a monitoring device) as well as efforts to comply with researcher or societal expectations (e.g., that PA is healthy and desirable), and is one of many possible research participation effects (McCambridge et al., 2014). When averaged across days, initial increases that may not represent 'typical' PA can inflate and skew estimates (Baumann et al., 2018; Clemes & Deans, 2012). This could lead to incorrect conclusions about which populations are already active (vs. not) and/or mask the effects of interventions for insufficiently active groups (French & Sutton, 2010).

Consequently, researchers are encouraged to consider potentially burdensome approaches such as adding days of observation to the start of a measurement period (which would be dropped from analyses; Clemes & Deans, 2012; Foote et al., 2017; König et al., 2022) and avoiding use of measurement devices that show recorded data to participants (French et al., 2021; König et al., 2022). These approaches

can increase study costs and participant burden, while the need for such procedures is not clear – particularly in populations that are regularly targeted for PA promotion. For example, conclusions about the presence and extent of measurement reactivity are clouded by differences in study type (e.g., observational vs. experimental) and inconsistencies in definitions and statistical approaches to modelling its effects (König et al., 2022).

Meta-analytic work does demonstrate that reactivity is stronger for PA than for other health behaviours (e.g., monitoring food intake through sensor devices; Bell et al., 2020; König et al., 2022). Importantly, however, evidence is equivocal as to the impact of measurement reactivity on PA estimates or conclusions, and attempts to induce PA measurement reactivity are largely unsuccessful (König et al., 2025). Some studies show no evidence of PA measurement reactivity (e.g., König et al., 2025; Ozdoba et al., 2004; Prewitt et al., 2013; Vincent & Pangrazi, 2002). Meta-analyses also indicate extensive variability in effect sizes: average effects are small to medium and are unlikely to meaningfully influence aggregated PA estimates (König et al., 2022). Thus, there appears to be considerable heterogeneity in PA measurement reactivity that could result from differences in research designs, or from differences in participant populations and characteristics. Further, existing work on PA measurement reactivity has focused on healthy populations (e.g., König et al., 2025) or children (e.g., Davis & Loprinzi, 2016; Zhu & Haegele, 2019). These groups are typically more active than the populations targeted for PA interventions and may show different responses to the introduction of PA monitoring (cf. Haynes & Robinson, 2019). Additional work is needed to understand the potential for reactivity to bias PA estimates in groups that regularly receive PA intervention, such as those with health conditions that increase their risk for cardiovascular disease (CVD; e.g., obesity, type 2 diabetes).

PA measurement reactivity among adults with CVD risk factors

CVD remains the leading cause of death worldwide (Karvinen et al., 2019), and adults in midlife (ages 40–60; Rodgers et al., 2019) are at particularly high risk of developing CVD and associated risk conditions that lead to premature mortality (e.g., hypertension; Matthews et al., 2009; Rodgers et al., 2019). Cardiovascular health in this group is negatively impacted by ageing of the cardiovascular system and psychosocial stressors (Cohen et al., 2015; Reeves et al., 2011; Sassarini, 2016). Insufficient PA in this group exacerbates these risks and represents a missed opportunity to reduce the likelihood of premature mortality – particularly given the additional and substantial benefits of modest PA engagement (e.g., stress reduction, improved quality of life; Chekroud et al., 2018; Kraus et al., 2019).

As a result, adults in midlife with CVD risk factors are consistently advised by healthcare providers to increase PA to improve cardiovascular health and longevity (Rodgers et al., 2019), and they represent the largest subset of participants in behavioural intervention programmes designed to increase PA (Cooke & Jones, 2017; Waters et al., 2011). This group often expresses interest in and invests considerable effort in participating in these interventions, though they experience only limited and/or short-term gains (Edney et al., 2019; Neve et al., 2010). Reactivity appears to be stronger for participants who see a need for changes to their health (vs. those who do not; Barta et al., 2012; König et al., 2022; Poulton et al., 2019), though existing observational studies show little evidence of reactivity among midlife adults with CVD risk conditions (Arigo & König, 2024; Maher et al., 2024). Further examination of PA measurement reactivity in this population across study designs is warranted, as the need for and effectiveness of PA promotion efforts for this at-risk group is contingent on accurate PA measurement.

Individual and study characteristics as moderators of reactivity

As noted, existing evidence indicates considerable heterogeneity in reactivity outcomes and effect sizes, and has not yet determined whether demographic, medical, psychological, or research design factors influence reactivity. For example, reactivity may differ by gender: relative to men, women demonstrate

greater responsiveness to and interest in others' perceptions of them (e.g., researchers' perceptions; Ambwani & Chmielewski, 2013; Tang et al., 2022). Individuals who are particularly prone to social comparison (i.e., self-evaluation relative to others; Arigo et al., 2021; Buunk & Ybema, 1997) and those who have made more (vs. fewer) prior efforts to improve their health (e.g., weight loss attempts; Barta et al., 2012) may also respond more strongly to the introduction of measurement, due to awareness of researcher observation and the desire to fulfil expectations (e.g., to live up to their peers in this respect). Conversely, medical and psychological factors such as body mass index (BMI) and depressive symptoms are negatively associated with PA levels; as those with higher (vs. lower) BMIs and depressive symptom severity are less active than others in their demographic groups (Hansen et al., 2013; Simon et al., 2008), their PA may be less responsive to environmental influences such as measurement procedures.

Several study design characteristics may also be relevant. First, differences in measurement devices (e.g., commercial vs. research-grade devices), location of the device (e.g., hip vs. wrist), and the availability of feedback from the device has the potential to impact PA behaviour (Clemes & Deans, 2012; French et al., 2021). In particular, feedback may increase the salience of PA measurement and prompt adjustments that would not occur under typical conditions. Second, the purpose of a research study may influence reactivity, such that participants who complete assessments prior to the start of an intervention and those who participate in an observation-only study may have different expectations for increases in future PA (Motl et al., 2012). Perhaps most importantly, definitions of reactivity are not consistent across studies. Some research defines reactivity as a linear decrease in PA over days of observation (e.g., Klenk et al., 2019), whereas others use comparisons between 'early' and 'later' days in the study (e.g., Hilgenkamp et al., 2012), and the definition of these periods is inconsistent. The lack of consensus for how reactivity is defined and evaluated causes confusion about its influence and about whether prevention efforts are warranted. Examining reactivity across multiple research contexts, with an emphasis on standardized effect sizes and impact on PA averages, would address critical questions about this phenomenon and refine evidence-based guidance for addressing it in PA research. Conducting this work among adults with CVD risk conditions – who show insufficient PA in daily life and are often the targets of PA intervention – provides the optimal overarching context for this work.

Aims of the present study

Toward these goals, we designed the present study to examine PA measurement reactivity across multiple datasets that included device-based PA assessment, among adults in midlife with one or more CVD risk conditions. We used a pre-specified, peer-reviewed coordinated secondary analysis approach, which was published as a protocol paper (Baga et al., 2025). A coordinated analysis approach leverages existing data resources and enables efficient comparison of analytic results and conclusions across these resources (Hofer & Piccinin, 2009), allowing for replication of substantive conclusions across samples and research contexts (cf. Hill et al., 2021; Mogle et al., 2021). This approach is particularly well suited to testing for PA measurement reactivity: if this phenomenon is indeed widespread (cf. French & Sutton, 2010), it should be observable across studies that use daily device-based assessment of PA and across PA outcomes of interest (e.g., activity counts, steps per day). Moreover, if PA measurement reactivity is problematic enough to warrant additional burdensome procedures, its effects should be not just statistically significant using *p*-values, but should affect PA estimates to an extent that reaches a threshold of minimal potential for meaningful impact (i.e., likelihood of meaningful differences in the real world).

In this spirit, the aims of the present study were to: (1) characterize PA measurement reactivity among adults ages 40–60 with CVD risk conditions following the introduction of measurement devices, (2) test for moderating effects of demographic, medical, and psychosocial characteristics on reactivity patterns, and (3) test for differences in reactivity based on study design and PA measurement features. Datasets included in the present analyses used either observation only or observation prior to the start of a behavioural intervention and were either publicly available or available by request to the principal investigator. As described

below, we set a minimum effect size threshold for detecting a potentially meaningful pattern of measurement reactivity in addition to our specification of statistical significance.

MATERIALS AND METHODS

Datasets were selected for inclusion based on availability of day-level PA data from individuals in the population of interest, across 6–7 days of measurement (i.e., the typical window of observation; Hilgenkamp et al., 2012). These data were from two publicly available databases under national long-term research programmes and four randomized clinical intervention trials in the United States, which collected observations between January 2005 and January 2023 (total included $N = 1825$). Publicly available datasets used observation-only research designs and included the National Health and Nutrition Examination Survey (NHANES; Centers for Disease Control, 2024) and the Midlife in the United States Study (MIDUS; Brim et al., 2019; combined $n = 1385$). NHANES captures health data in a national sample of US adults and provides these data for public use online; data for the current study were from the 2013–2014 phase of data collection, as this was the most recent phase to release PA data (7 days of assessment). MIDUS assesses health-related factors among US adults ages 25–75; data included in the present analyses were collected during the MIDUS Biomarker Project (6 days of assessment, 2004–2009; Ryff et al., 2010). NHANES and MIDUS captured PA behaviour using ActiGraph GT3X+ and Actiwatch-64 devices, respectively. Both datasets are publicly available through the International Consortium for Political and Social Research repository hosted by the University of Michigan (ICPSR; United States). Participants in these studies had not enrolled in a previous wave of data collection (i.e., were new to study-specific measurement procedures).

Clinical trials included four tests of behavioural weight loss treatments that were funded by the US National Institutes of Health. These studies included Project ENACT, Project IMPACT, FitLink Pilot, and FitLink Full (combined $n = 440$). Project ENACT (NCT01858714; completed 2011–2012) and Project IMPACT (NCT02363010; completed 2014–2016) examined the effects of specific enhancements to behavioural weight loss treatment on long-term weight loss maintenance (i.e., incorporation of acceptance-based skills, greater emphasis on PA). These studies both provided in-person treatment with 7 days of PA assessment at baseline (prior to treatment start), using ActiGraph GT3X+ devices that were worn on the hip. The FitLink Pilot study (NCT03337139; completed 2018–2019) and FitLink Full study (2021–2023; NCT05180448) evaluated the effects of sharing self-monitoring data (e.g., weight loss, PA behaviour) with an individual health coach, with treatment group members, and/or with a nominated friend or family member on long-term weight loss maintenance. Both FitLink studies assessed PA for 7 days at baseline using wrist-worn Fitbit devices; the Pilot provided in-person treatment while the Full version was conducted remotely. Additional details are provided in Table 1 and in Baga et al. (2025).

Participants

Individuals eligible for inclusion in analyses were adults who provided data for the aforementioned studies who (1) were ages 40–60 (inclusive) at the time of participation, and (2) reported or showed evidence of ≥1 health condition(s) that confer CVD risk at the time of enrollment. CVD risk conditions were type 2 diabetes, prediabetes, high cholesterol, hypertension, smoking tobacco, and obesity (i.e., BMI ≥ 30 kg/m^2). Each study assessed height and weight, as described below, though each assessed only a subset of the other conditions considered for inclusion. Further, participants were included only if they provide adequate PA data to support planned analyses, with the following criteria: (1) device- assessed PA data on day 1 of the observation period, (2) ≥3 consecutive days of device-assessed PA data (total), and (3) ≥10 h of device wear time where this indicator was available (i.e., for NHANES, MIDUS, ENACT, IMPACT). Day-level wear time was not available for Fitbit data; participants were included in each parent clinical trials dataset if they had ≥6 days of Fitbit data at baseline (i.e., recorded a minimum

TABLE 1 Measurement of physical activity and individual differences across six studies.

	NHANES[a] ($n=1208$)	MIDUS ($n=177$)	ENACT ($n=143$)	IMPACT ($n=140$)	FitLink pilot ($n=27$)[b]	FitLink full ($n=130$)
PA device	ActiGraph GT3X+	Actiwatch-64	ActiGraph GT3X+	ActiGraph GT3X+	Fitbit	Fitbit
Wear location	Wrist	Wrist	Hip	Hip	Wrist	Wrist
Duration	7 days	6 days	7 days	7 days	7 days	7 days
Outcomes	MIMS units	TAC	Steps	Steps	Steps	Steps
BMI	Measured	Measured	Measured	Measured	Measured	Measured
CVD Risk[a,c]	Hypertension	Hypertension	Hypertension	Hypertension	Hypertension	Hypertension
	Diabetes	Diabetes	Diabetes	Diabetes	Diabetes	Diabetes
	Obesity	Obesity	Obesity	Obesity	Obesity	Obesity
	Smoker	Smoker	Smoker	Smoker	Smoker	
	High cholesterol					
	Prediabetes					
Depressive symptoms	PHQ-9	CES-D	BDI-II	BDI-II	WALI	BDI-II
PA Motivation			TSRQ	TSRQ	TSRQ	
Weight loss attempts	Self-reported (past year)		WALI	WALI	WALI	WALI
Social comparison		Comparison of CVD risk	INCOM			INCOM

[a]CVD risk questions were captured using individual questions about these conditions and factors for observational studies and as part of the Weight and Lifestyle Inventory for clinical trials.

[b]Sample size removed four additional participants than originally intended in protocol (see Baga et al., 2025), due to focus only on the first week of PA data.

[c]BMI was calculated based on measured height and weight.

of 500 steps per day on 6 days) and were included in the present analyses if they met all other criteria above. Participant demographics for each study are available in Table 2; there were no significant demographic differences between included and excluded participants in any study.

Measures

Demographics

These data were captured at baseline using surveys or interviews and included age, gender, income, education, marital status, and racial/ethnic identity. NHANES and MIDUS collected this information through interviews; all four clinical trials used electronic surveys.

Physical activity (PA) behaviour

PA was measured using research-grade or commercially available PA monitors to generate daily summary metrics for PA. Publicly available datasets used wrist-worn ActiGraph devices (ActiGraph GT3X+, ActiGraph-64) and measured PA behaviour using Monitor Independent Movement Summary units (MIMS, NHANES; John et al., 2019) and total activity counts (TAC, MIDUS; Ryff et al., 2010). Clinical trials used hip-worn ActiGraph devices (ENACT, IMPACT) or wrist-worn Fitbit devices (FitLink Pilot: Fitbit Flex; FitLink Full: Fitbit Inspire 2) to monitor steps per day during a 7-day baseline period prior to the start of the behavioural weight loss intervention (see Table 1).

CVD risk conditions

Participants' CVD risk status was evaluated using baseline questionnaires or physical assessment. All participants were assessed for the presence of hypertension, type 2 diabetes, and obesity (i.e., BMI $\geq 30\,kg/m^2$ captured through measurement or self-report). Additional risk factors that informed eligibility for a subset of studies included self-reported high cholesterol, prediabetes, and current smoker (see Table 1 for conditions assessed in each study).

Depressive symptoms

Observational studies captured these symptoms using the Patient Health Questionnaire (PHQ-9; Kroenke & Spitzer, 2002). Clinical trials used the Beck Depression Inventory (BDI-II; Beck et al., 1996) or items assessing depressed mood and anhedonia from the Weight and Lifestyle Inventory (FitLink Pilot; WALI; Wadden & Foster, 2006). Total scores for these measures range from 0 to 27 (PHQ-9) and 0 to 63 (BDI-II), respectively. Items assessing depressed mood and anhedonia in the WALI were dichotomized and summed to capture presence and severity of these symptoms among FitLink Pilot participants. In all cases, higher scores indicate more severe depressive symptoms.

Social comparison

This construct, which captures participants' experiences with self-evaluation relative to others, was measured for three studies (MIDUS, ENACT, FitLink Full). Assessment for MIDUS focused on health-specific social comparison and asked for participants' perception of their risk of having a heart

TABLE 2 Demographics for included participants (mean ± sd or *n* (%)).

	NHANES[a] (*n* = 1208)	MIDUS (*n* = 177)	ENACT (*n* = 143)	IMPACT (*n* = 140)	FitLink pilot (*n* = 27)	FitLink full (*n* = 130)
Age	49.9 ± 6.0	50.9 ± 5.8	52.3 ± 5.6	52.9 ± 5.0	50.2 ± 6.1	51.5 ± 5.7
Gender						
Women	628 (52.0)	106 (59.3)	117 (81.8)	116 (82.9)	25 (92.6)	109 (83.8)
Men	580 (48.0)	72 (40.7)	26 (18.2)	24 (17.1)	2 (7.4)	21 (16.2)
Race[a]						
AI/NA[b]	–	0 (.0)	1 (.7)	0 (.0)	1 (3.2)	0 (.0)
Asian	119 (9.8)	2 (1.1)	1 (.7)	1 (.7)	0 (.0)	3 (2.3)
NH or OPI[b]	–	0 (.0)	1 (.7)	0 (.0)	0 (.0)	0 (.0)
Black/AA[b]	283 (23.4)	50 (28.4)	51 (35.7)	35 (25.0)	12 (38.7)	17 (13.1)
White	492 (40.7)	103 (58.5)	84 (58.7)	98 (70.0)	14 (45.2)	107 (82.3)
Other/mixed race	34 (2.8)	21 (11.9)	5 (3.5)	6 (4.3)	4 (12.9)	3 (2.3)
Ethnicity[a]						
Hispanic or Latino	280 (23.2)	0 (.0)	8 (5.6)	5 (3.6)	2 (7.4)	10 (7.7)
Not Hispanic or Latino/a	946 (77.7)	177 (100.0)	134 (94.4)	135 (96.4)	25 (92.6)	120 (92.3)
Marital status						
Married	721 (59.2)	100 (56.5)	86 (59.3)	91 (65.0)	12 (44.4)	90 (68.7)
Widowed	28 (2.3)	4 (2.3)	2 (1.4)	2 (1.4)	1 (3.7)	1 (.8)
Divorced	203 (16.6)	32 (18.1)	22 (15.2)	21 (15.0)	3 (11.1)	6 (6.2)
Separated	52 (4.3)	9 (5.1)	3 (2.1)	5 (3.6)	3 (11.1)	–
Never married	151 (12.4)	32 (18.1)	32 (22.1)	21 (15.0)	8 (29.6)	–
Single	–	–	–	–	–	20 (15.3)
Cohabitating	63 (5.2)	–	–	–	–	9 (6.9)
Not cohabitating	–	13 (16.9)	–	–	–	2 (1.5)
Income[a]						
$0 to $25k	294 (25.5)	63 (36.6)	9 (6.4)	4 (2.9)	1 (3.7)	–
$25k–$50k	210 (18.2)	53 (30.8)	21 (14.9)	13 (9.5)	4 (14.8)	–

TABLE 2 (Continued)

	NHANES[a] (*n* = 1208)	MIDUS (*n* = 177)	ENACT (*n* = 143)	IMPACT (*n* = 140)	FitLink pilot (*n* = 27)	FitLink full (*n* = 130)
$45k-$55k[a]	92 (7.9)	–	–	–	–	–
$50k–$75k	122 (10.6)	27 (15.7)	31 (22.0)	15 (11.0)	5 (18.5)	–
$75k–$100k	113 (9.8)	14 (8.1)	24 (17.0)	23 (16.8)	6 (22.2)	–
>$100k	251 (21.8)	8 (4.6)	13 (9.2)	20 (14.6)	2 (7.4)	–
$125k–$150k	–	2 (1.2)	18 (12.8)	17 (12.4)	4 (14.8)	–
$150k–$175k	–	2 (1.2)	8 (5.7)	18 (13.1)	0 (.0)	–
$175k–$200k	–	0 (.0)	5 (3.6)	13 (9.5)	2 (7.4)	–
>$200k	–	3 (1.7)	12 (8.5)	14 (10.2)	3 (11.1)	–
Education						
Less than 9th grade	80 (6.8)	2 (1.1)	0 (.0)	0 (.0)	0 (.0)	–
Partial high school	156 (13.4)	10 (5.6)	1 (.7)	0 (.0)	0 (.0)	–
High School or GED	280 (24.0)	35 (19.8)	13 (9.1)	5 (4.3)	1 (3.7)	–
Associate's degree, technical degree, or partial college	373 (31.9)	52 (43.5)	22 (15.4)	19 (16.4)	3 (11.1)	–
Bachelor's degree	279 (23.9)	44 (24.9)	52 (36.4)	41 (35.3)	10 (37.0)	–
Graduate or professional degree	–	34 (19.2)	55 (38.5)	51 (44.0)	13 (48.2)	–
BMI (kg/m^2)	31.2 ± 7.4	32.0 ± 7.2	36.3 ± 4.5	35.5 ± 4.4	36.7 ± 4.6	36.5 ± 4.8
BMI category						
<18.5 kg/m^2	9 (.8)	1 (.6)	0 (.0)	0 (.0)	0 (.0)	0 (.0)
18.5–25 kg/m^2	224 (18.6)	26 (14.7)	0 (.0)	0 (.0)	0 (.0)	0 (.0)
25–30 kg/m^2	321 (26.6)	43 (24.3)	7 (4.9)	6 (4.4)	0 (.0)	4 (3.1)
>30 kg/m^2	653 (54.1)	107 (60.4)	136 (95.1)	131 (95.6)	27 (100.0)	126 (96.9)

[a]NHANES demographics information: Ethnicity and race were asked using a single question, combining both characteristics. Income was assessed for the household (rather than the individual); Ethnicity and race were asked using a single question, combining these characteristics.

[b]Race categories: AI/NA – American Indian/Native Alaskan; NH or OPI – Native Hawaiian or Other Pacific Islander; AA – African American.

attack relative to others of their same gender and age (i.e., higher, lower, the same). The other studies used the Iowa–Netherlands Comparison Orientation Measure (INCOM; Gibbons & Buunk, 1999) to evaluate tendencies towards making social comparisons on a continuous scale. Participants rated items such as *I often compare myself with others with respect to what I have accomplished in life* on a scale from 1 (*Disagree strongly*) to 5 (*Agree strongly*). This measure provides a score for overall tendencies towards social comparisons as well as scores for tendencies towards upward (i.e., 'better off') and downward (i.e., 'worse off') comparison targets. Higher scores indicate stronger inclination to compare.

PA motivation

Three clinical trials (ENACT, IMPACT, FitLink Pilot) measured PA motivation using the Treatment Self-Regulation Questionnaire (TSRQ; Levesque et al., 2007). This measure provides separate scores for specific types of motivation (i.e., autonomous motivation, introjected regulation, external regulation, amotivation). Participants rate items such as *I want to be physically active because I personally believe it is the best thing for my health* (autonomous motivation) on a scale from 1 (*Not at all*) to 7 (*Very true*), with higher scores representing greater motivation of that subtype.

Weight loss attempts

The number of prior weight loss attempts was measured for NHANES and the four clinical trials datasets. Participants were asked to report attempts in the past year (NHANES) or all past attempts (as part of the WALI in clinical trials; Wadden & Foster, 2006). These efforts were summed to capture total weight loss attempts in the relevant time frame.

Procedures

This study was approved as exempt by the Institutional Review Board at Rowan University and procedures were described in a published, peer-reviewed protocol paper prior to conducting analyses (Baga et al., 2025). Datasets and relevant documentation were accessed online for publicly available data (via ICPSR) and were shared by the home institution for randomized clinical trials. Data acquired and prepared for planned analyses included the demographic information, CVD risk factors, scores for individual difference measures, and PA-relevant behavioural parameters that were required for assessment of eligibility and examination of PA (e.g., number of days with valid observations). CVD risk variables were coded dichotomously based on presence (1) or absence (0) of each condition. As studies assessed different combinations and numbers of CVD risk factors, dichotomous variables were used to calculate the proportion of CVD risk factors for each person out of the total number of CVD risk conditions assessed in the relevant study.

Next, a dichotomous eligibility variable was created to indicate whether a participant met the age and CVD risk criteria for planned analyses (i.e., 1 vs. 0; analyses included only those coded as 1). PA behaviour was analysed at the day level, along with the date of observation (where available) and sequential day of observation for each summary measure of PA for that day. Variables were created to indicate the number of continuous measurement days per participant and whether the measurement day occurred on a weekday or weekend (where available). Inclusion in analyses was further determined by identifying participants with ≥3 days of PA data in the relevant unit (i.e., MIMS, TAC, steps per day), including the first measurement day, and days with ≥10 h of device wear time (where available).

Data analysis

As described above, participants were identified for inclusion if they met demographic and health-related criteria (i.e., ages 40–60, ≥1 CVD risk factors) as well as PA device wear time criteria. PA measurement criteria were applied following exclusion based on age and CVD risk status, which removed 11 additional participants from NHANES, 16 participants from FitLink Pilot, and 21 participants from FitLink Full; none were removed from MIDUS or the other clinical trials datasets. The final analytical sample sizes are presented in Table 2; as noted, there were no differences between those included and excluded from analyses. The final sample included 11,569 valid PA observations across 1825 adults. With the exception of the FitLink Pilot, each dataset afforded power >.80 to detect linear (main) effects of time at an effect size equivalent to $sr \geq .25$ (see further detail below; Hoffman, 2015; Hox et al., 2017). Power to detect moderated effects was more limited in most datasets; for this reason, and to enable the intended comparisons across datasets, we relied on effect sizes to determine the potential impact of reactivity patterns on means and conclusions; see below for details about the inference criteria.

We used multilevel modelling to account for nesting, with days of observation (level 1) nested in participants (level 2). Analyses were completed in SAS 9.4 (Cary, NC) using PROC MIXED with restricted maximum likelihood estimation. In line with a coordinated analysis method (Hill et al., 2021; Hofer & Piccinin, 2009), each dataset was treated separately rather than pooling across datasets; this approach carries advantages with respect to addressing the potential generalizability of an effect and allowing for examination of differences between studies (Graham et al., 2022). In each dataset (separately), we used empty models to identify (1) intraclass correlation coefficients (ICCs) to estimate between-person stability versus within-person variability in PA outcomes across days, and (2) averages for each PA outcome across days in each study (for comparison with published norms). As noted, PA outcomes available were different across studies: NHANES used MIMS units, MIDUS used TAC, and clinical trials used steps per day. We examined the influence of potential covariates in each dataset, including age, day of week (weekday vs. weekend) and BMI on PA outcomes (cf. Hansen et al., 2013; Klenk et al., 2019). Day of week and BMI were significantly associated with PA in three out of six datasets; BMI was included as a covariate in analyses for all datasets, and day of week was included where it was available (i.e., all datasets other than MIDUS; see Table 3). BMI and continuous moderators were grand-mean centred in all models. Of note, the statistical patterns and primary conclusions did not differ with these covariates removed from analyses.

Our first aim was to characterize the extent of PA measurement reactivity in each dataset; we did this by modelling the fixed linear effect of day in study on PA outcomes using measurement day as a continuous predictor (cf. Klenk et al., 2019). Day in study was also examined as a categorical predictor, comparing Day 1 to all other days and days 1–2 to all other days using planned contrasts (cf. Hilgenkamp et al., 2012). Given that there is no consistent method for identifying PA measurement reactivity, we expected to observe either linear trends, differences in planned contrasts, or both. Effect sizes were calculated in two ways: (1) the difference in the relevant PA unit for each outcome between days (e.g., steps per day), and (2) using semipartial correlation coefficients as a standardized estimate, which enabled direct comparison across datasets. As there is also no agreed-upon effect size estimate for multilevel models (cf. Lorah, 2018), to promote interpretability across fields that conduct device-based assessments of PA, we also report the corresponding Cohen's *d* value with 90% confidence intervals and used these to inform our interpretations (described further below).

We addressed our second aim of identifying individual differences in PA measurement reactivity patterns in two ways. We first added random effects to determine whether this improved model fit; significant improvement would indicate systematic variability in the level 1 effect of interest at level 2 (i.e., between people). Model comparisons were based on χ^2 tests of the −2 log likelihood difference between the full and reduced models. Second, and independent of whether random effects improved model fit, we tested for moderation of linear and categorical time effects by the individual difference characteristics of interest using cross-level interactions. These included gender, BMI, number of CVD risk factors,

TABLE 3 Main effect model results for each study.

	NHANES *B* (*SE*)	MIDUS *B* (*SE*)	ENACT *B* (*SE*)	IMPACT *B* (*SE*)	FitLink pilot *B* (*SE*)	FitLink full *B* (*SE*)
ICC	.66	.62	.49	.39	.36	.35
Intercept	15.10 (.12)**	339.32 (9.30)**	6278.22 (295.56)**	6395.49 (256.81)**	7168.22 (859.45)**	7124.62 (339.40)**
BMI (kg/m^2)	−.08 (.01)**	−2.25 (1.09)	12.27 (46.40)	−158.10 (41.56)*	−109.56 (122.81)	−106.87 (42.48)
Day of week (weekday vs. weekend)	.17 (.06)*	–	115.47 (190.45)	809.32 (188.09)**	11.32 (539.07)	−322.78 (217.04)
Day of observation	−.04 (.01) *sr* = .10 *d* = −.09 (−.13, −.04)	−1.18 (1.45), *sr* = .06 *d* = −.06 (−.17, .06)	−62.57 (42.25) *sr* = .12 *d* = −.12 (−.27, .04)	−52.81 (43.70) *sr* = .10 *d* = −.09 (−.25, .01)	182.21 (126.12) *sr* = .26[a] *d* = .23 (−.03, .50[a])	−58.55 (48.93) *sr* = .10 *d* = −.09 (−.21, .03)
Day 1 vs. Other days	−.04 (.01)*	−1.18 (1.45)	−62.57 (42.25)	−52.81 (43.70)	182.21 (126.12)	−58.55 (48.93)
Days 1 & 2 vs. Other days	−.02 (.007)*	−.59 (.73)	−31.28 (21.12)	−26.40 (21.85)	91.12 (63.06)	−29.28 (24.47)

Note: PA outcome for NHANES: MIMS/min, for MIDUS: activity counts/min; for ENACT, IMPACT, FitLink Pilot, and FitLink Full: steps/day; Standard effect sizes (*sr*) and Cohen's *d* (90% CI) were calculated to test significance.

[a]*sr* ≥ .25 or *d* upper-limit CI ≥ .20.

*$p < .01$; **$p < .0001$.

severity of depressive symptoms, social comparison responses, PA motivation, and weight loss history. Given prior work in these areas, we expected PA measurement reactivity to be more pronounced among women; those with greater (vs. lesser) PA motivation and social comparison responses; those with lower (vs. higher) BMIs, depressive symptom scores, and numbers of CVD risk factors; and those with more (vs. fewer) weight loss attempts. We also tested for gender differences in moderated effects, which were expected to be stronger among women, using three-way interactions. We conducted separate models for each three-way interaction, and each model also included the main effects and corresponding two-way interactions. Addressing our third aim involved comparison of reactivity patterns and effect sizes between studies that used different PA measurement methods. We compared study designs (i.e., observation vs. pre-intervention) and device types (i.e., worn on wrist vs. hip, offered feedback vs. no feedback).

The threshold for statistical significance was set at $p < .05$ for two-tailed tests. Effect sizes are initially expressed as semipartial correlation coefficients for comparison across datasets, which are appropriate for multilevel models. Semipartial correlation coefficients (*sr*s) are comparable to Cohen's *d* and are interpreted similarly with respect to size (i.e., .20 as a small effect, .50 as a moderate effect, .80 as a large effect; Cohen, 1988). However, *d* is typically used to describe group-level, between-person differences, whereas *sr*s in this study describe day-level, within-person differences (or linear change across days). We note that there is no standardized effect size that automatically equates to clinical or practical significance; however, intervention science consistently relies on effect size interpretation as suggesting small, medium, and large effects in the real world (Rutledge & Loh, 2004). To determine the potential for *minimally meaningful potential impact* of PA measurement reactivity patterns at the day level, we set $sr \geq .25$; this is roughly equivalent to Cohen's *d* of .20, or a small but meaningful effect at the day level (cf. Ferguson, 2009; Panjeh et al., 2023).

To supplement our planned analyses as specified in Baga et al. (2025), we also conducted equivalence tests based on Cohen's *d*; these tests were one-tailed, as we were primarily interested in evidence of initial elevation effects (given that these would align with prior findings). Upper bounds of 90% confidence intervals for $d \geq .20$ were interpreted as rejecting the null (i.e., null: the effect is larger than the specified size). As noted, where available, we also report day-level differences and trends in steps per day, to provide additional estimates with respect to real-world outcomes. Both the inclusion of Cohen's *d* with confidence intervals and equivalence tests were conducted as deviations (i.e., supplements) to the prespecified analysis plan, in response to helpful suggestions from reviewers.

RESULTS

Relative to published estimates of PA behaviour among adults ages 40–60 with risk factors for CVD, participants in the studies included in this analysis engaged in similar levels of PA (i.e., moderately active but not meeting US PA guidelines; Aguiar et al., 2024; Kraus et al., 2019). Specifically, participants in NHANES and MIDUS achieved an average of 15.0 MIMs per minute ($SE = .09$) and 335.2 activity counts per minute ($SE = 7.91$) across days, respectively. Participants in clinical trials completed an average of 6123.75–7883.32 steps per day prior to the start of the intervention. ICCs for each study ranged from .35 to .66 (see Table 3), indicating that 34%–65% of the variability in each PA outcome could be attributed to within-person variability (and error) across days of observation.

Aim 1: Extent of PA measurement reactivity

Fixed linear effects generally showed non-significant downward trends in PA behaviour across study days (ps >.14, *sr*s = .06–.12, *d*s = −.12 to −.06) and planned contrasts between early and later days were not significant (ps >.14; see Table 3 and Figure 1). In subsequent equivalence tests, the upper bound of 90% confidence intervals for associated effect sizes did not reach the specified threshold of $d \geq .20$, suggesting that the effect is smaller than this threshold (and thus, the null that the effect size is larger than this threshold should be rejected; see Table 3). NHANES was an exception, for which the negative

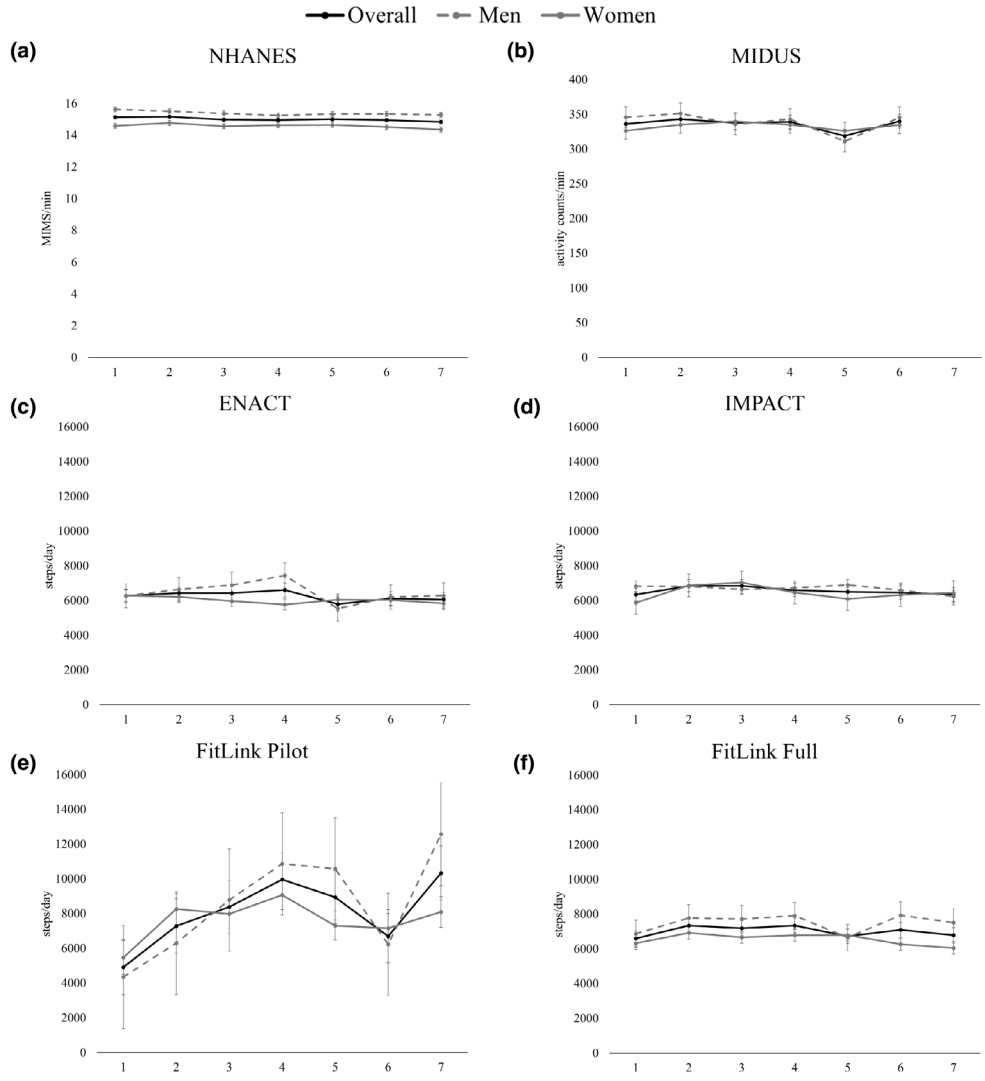

FIGURE 1 Physical activity outcomes across observation days, overall and by gender, in each included dataset. MIMs, monitor independent movement summary units.

linear trend and contrasts between early and later days were statistically significant ($F[16432] = 11.88$, $ps = .0006$). Importantly, however, the effect sizes for these models were $sr = .10$ and contrasts showed differences of only .02–.04 MIMS units; as these effects were equivalent to $d = -.09$ and the upper bounds of confidence intervals were <.20, it is unlikely that reactivity would meaningfully affect PA estimates. In addition, although not statistically significant ($ps = .15$), the FitLink Pilot showed an effect size of $sr = .26$ (or $d = .23$, 90% CI: −.03, .50) and potentially meaningful difference between days: Day 1 was 91 steps higher than the average of subsequent days and the average of days 1–2 was 182 steps higher than the average of subsequent days.

Aim 2: Individual differences in PA measurement reactivity

Adding random effects resulted in model non-convergence for MIDUS and ENACT and did not improve model fit for most of the included datasets ($\chi^2 = .70–8.80$, $ps > .05$; see Table 4). Adding

TABLE 4 Random effect model results for each study.

	NHANES *B* (*SE*)	MIDUS *B* (*SE*)	ENACT *B* (*SE*)	IMPACT *B* (*SE*)	FitLink pilot *B* (*SE*)	FitLink full *B* (*SE*)
Intercept	12.93 (.67)*	9774.37 (1161.53)**	5,095,427 (751979)**	3,966,633 (826695)**	4,848,534 (3781718)	6,797,717 (1564456)*
Slope	.12 (.01)*	–	–	103,775 (43760)*	39,552 (137261)	106,958 (50277)
Intercept	15.11 (.13)*	339.32 (9.30)**	6393.69 (269.88)**	6414.21 (259.66)**	7193.96 (818.63)**	7171.80 (362.33)*
BMI (kg/m^2)	−.08 (.01)*	−2.24 (1.02)	12.27 (46.40)	−161.64 (41.54)*	−107.25 (121.10)	−105.87 (42.35)
Day of week (weekday vs. weekend)	.14 (.06)	–	−115.47 (190.45)	774.21 (187.32)**	23.40 (536.36)	−373.72 (214.00)
Day of observation	−.04 (.02)	−1.18 (1.45)	−62.57 (42.25)	−48.94 (51.13)	175.66 (130.88)	−61.25 (55.50)
χ^{2a}	150.4**	–	–	8.8	.7	6.8

Note: PA outcome for NHANES: MIMS/min, for MIDUS: activity counts/min; for ENACT, IMPACT, FitLink Pilot, and FitLink Full: steps/day.

[a]Compared to main effect model results (see Table 3).

*$p < .01$; **$p < .0001$.

random effects did improve NHANES model fit ($\chi^2 = 150.4$, $p < .001$), however. There were no differences in reactivity patterns between genders in most studies (ps >.08, srs = .02 to .09, $d = -.08$ to .01 and upper bounds of 90% confidence intervals for d did not reach the specified threshold of $d \geq .20$; see Table 5 and Figure 1). The FitLink Pilot was an exception: though not statistically significant ($p = .08$), there was a potentially meaningful difference in PA reactivity by gender ($sr = .31$; $d = -.28$, 90% CI: −.55, −.02). Men tended to show a greater linear *increase* in steps per day over study days as compared with women (~922 steps/day vs. 119 steps/day, respectively), which represents a pattern opposite of what typically characterizes reactivity. As expected, BMI, number of CVD risk factors, and depressive symptom severity all were negatively associated with PA behaviour across days (i.e., person-level averages), such that individuals with higher BMIs, more CVD risk factors, and more severe depressive symptoms engaged in less PA on average (across days) than their counterparts. However, statistically significant main effects of these characteristics showed up only in NHANES (BMI, CVD risk) and IMPACT (BMI) and these did not interact with day in study (ps >.13, srs = .01–.14; ds = .0 to −.06; see Table 6). PA motivation, weight loss attempts, and social comparison also did not moderate these effects.

When incorporating the three-way interaction between gender and other moderators, two notable differences emerged. In MIDUS, more severe depression was associated with a decrease in activity counts over study days, and this effect was attenuated in men (relative to women; $p = .007$). The size for this effect was $sr = .20$ and the difference between women with higher (vs. lower) depression scores was only 1 activity count per minute per day. As the upper bound of the 90% confidence interval exceeded $d = .20$ in this case (Cohen's $d = .18$, 90% CI: .07, .30), there remained potential for measurement reactivity among women with higher (vs. lower) depression scores. In contrast, there was a statistically significant *and* moderate effect size in FitLink Full for the three-way interaction between study day, gender, and social comparison ($p = .003$, $sr = .26$, $d = -.22$, 90% CI: −.34, −.10). For women, a stronger (vs. weaker) social comparison tendency was associated with an *increase* in steps per day over study days; this effect was the opposite for men, who experienced a reduction of ~80 steps per day.

TABLE 5 Gender moderation effect model results for each study.

	NHANES *B* (*SE*)	MIDUS *B* (*SE*)	ENACT *B* (*SE*)	IMPACT *B* (*SE*)	FitLink pilot *B* (*SE*)	FitLink full *B* (*SE*)
Intercept	15.51 (.15)*	331.88 (12.13)**	6657.26 (652.65)**	6060.57 (542.64)**	4974.78 (2739.23)	7521.03 (715.70)**
BMI (kg/m^2)	−.09 (.01)*	−2.12 (1.11)	16.36 (46.88)	−157.76 (41.70)*	−109.89 (124.76)	−100.49 (42.36)
Day of week (weekday vs. weekend)	.17 (.06)	–	112.35 (190.63)	809.77 (188.18)**	36.85 (535.86)	−327.63 (217.17)
Day of observation	−.05 (.02)	.29 (1.89)	−74.67 (101.57)	−8.31 (101.62)	921.64 (441.03)	−37.76 (119.37)
Gender (ref: female)	.86 (.21)*	18.44 (19.19)	456.82 (699.87)	404.12 (576.07)	−2373.78 (2813.64)	466.84 (757.23)
Day of observation × Gender	−.01 (.03) *sr* = .02 *d* = .01 (−.03, .05)	−3.60 (2.96) *sr* = .09 *d* = −.08 (−.20, .03)	14.65 (111.68) *sr* = .01 *d* = .01 (−.12, .14)	−54.24 (112.57) *sr* = .04 *d* = −.04 (−.16, .09)	−803.02 (459.72) *sr* = .31[a] *d* = −.28 (−.55, −.02)	115.38 (130.29) *sr* = .08 *d* = −.06 (−.18, .05)
Day 1 vs. Other days	.01 (.03)	−3.60 (2.96)	14.65 (111.68)	−54.24 (112.57)	−803.02 (459.72)	−115.38 (130.29)
Days 1 & 2 vs. Other days	.007 (.01)	−1.80 (1.48)	7.33 (55.84)	−27.12 (56.29)	−401.51 (229.86)	−57.69 (65.15)

Note: PA outcome for NHANES: MIMS/min, for MIDUS: activity counts/min; for ENACT, IMPACT, FitLink Pilot, and FitLink Full: steps/day; Standard effect sizes (*sr*) and Cohen's *d* (90% CI) were calculated to test significance.

[a]*sr* ≥ .25 or *d* upper-limit CI ≥ .20.

*$p < .01$; **$p < .0001$.

TABLE 6 Other moderation effect model results for each study.

	NHANES *B* (*SE*)	MIDUS *B* (*SE*)	ENACT *B* (*SE*)	IMPACT *B* (*SE*)	FitLink pilot *B* (*SE*)	FitLink full *B* (*SE*)
BMI						
Day of observation	−.04 (.01)*	−1.19 (1.45)	−62.33 (42.25)	−51.50 (43.72)	182.41 (126.80)	−58.29 (48.97)
BMI	−.08 (.01)*	−1.50 (1.30)	56.76 (58.96)	−181.24 (49.52)*	−105.85 (168.06)	−117.76 (58.17)
Day of observation × BMI	.0009 (.002) *sr* = .02 *d* = .01 (−.03, .05)	−.23 (.21) *sr* = .08 *d* = −.07 (−.19, .04)	−11.46 (9.38) *sr* = .10 *d* = −.10 (−.23, .03)	8.84 (10.25) *sr* = .07 *d* = .06 (−.06, .19)	−.90 (28.07) *sr* = .00 *d* = −.00458 (−.27, .26[a])	2.77 (10.09) *sr* = .02 *d* = .02 (−.10, .14)
Day 1 vs. Other days	.0009 (.002)	−.23 (.21)	−11.46 (9.38)	8.84 (10.25)	−.90 (28.07)	2.77 (10.09)
Days 1 & 2 vs. Other days	.0005 (.0009)	−.11 (.10)	−5.73 (4.69)	4.42 (5.13)	−.45 (14.03)	1.38 (5.04)
CVD risk						
Day of observation	−.04 (.01)*	−1.18 (1.45)	−62.35 (42.28)	−52.74 (43.72)	173.02 (127.04)	−58.50 (48.96)
CVD risk	−.50 (.11)**	8.71 (10.66)	4.65 (423.22)	−234.69 (412.74)	−938.18 (1257.75)	−376.97 (502.44)
Day of observation × CVD risk	.01 (.01) *sr* = .03 *d* = .02 (−.02, .06)	.07 (1.64) *sr* = .00 *d* = .00275 (−.11, .12)	−38.63 (66.51) *sr* = .05 *d* = −.05 (−.18, .08)	−11.32 (85.74) *sr* = .01 *d* = −.00960 (−.13, .11)	−162.86 (204.95) *sr* = .14 *d* = −.13 (−.39, .14)	36.09 (85.90) *sr* = .04 *d* = .03 (−.09, .15)
Day 1 vs. Other days	.01 (.01)	.07 (1.64)	−38.63 (66.51)	−11.32 (85.74)	−162.86 (204.95)	36.09 (85.90)
Days 1 & 2 vs. Other days	.006 (.007)	.03 (.82)	−19.32 (33.25)	−5.66 (42.87)	−81.43 (102.48)	18.05 (42.95)
Depressive symptoms						
Day of observation	−.04 (.01)*	−1.03 (1.46)	−61.63 (42.23)	−55.70 (44.02)	183.15 (126.57)	−58.55 (48.96)
Depressive symptoms	−.06 (.02)	.55 (1.08)	−83.92 (40.56)	20.05 (33.06)	−2425.07 (1383.57)	−12.07 (74.24)

(Continues)

TABLE 6 (Continued)

	NHANES *B* (*SE*)	MIDUS *B* (*SE*)	ENACT *B* (*SE*)	IMPACT *B* (*SE*)	FitLink pilot *B* (*SE*)	FitLink full *B* (*SE*)
Day of observation × Depressive symptoms	.003 (.003) *sr* = .03 *d* = .02 (−.02, .07)	−.04 (.17) *sr* = .02 *d* = −.01 (−.13, .10)	9.79 (6.49) *sr* = .13 *d* = .12 (−.01, .25)	−9.02 (6.86) *sr* = .11 *d* = −.10 (−.22, .02)	33.50 (238.39) *sr* = .02 *d* = .02 (−.24, .29)	5.31 (12.77) *sr* = .04 *d* = .03 (−.09, .15)
Day 1 vs. Other days	.003 (.003)	−.04 (.17)	9.79 (6.49)	−9.02 (6.86)	33.50 (238.39)	5.31 (12.77)
Days 1 & 2 vs. Other days	.001 (.001)	−.02 (.09)	4.89 (3.24)	−4.51 (3.43)	16.75 (119.20)	2.66 (6.39)
PA motivation						
Day of observation	–	–	−60.02 (42.25)	−50.56 (44.70)	185.25 (126.38)	–
PA motivation	–	–	165.42 (86.86)	94.28 (79.80)	96.72 (176.94)	–
Day of observation × PA motivation	–	–	5.19 (14.24) *sr* = .03 *d* = .03 (−.10, .16)	−2.00 (16.71) *sr* = .01 *d* = .00899 (−.13, .11)	20.82 (29.57) *sr* = .13 *d* = .11 (−.15, .38)	–
Day 1 vs. Other days	–	–	5.19 (14.24)	−2.00 (16.71)	20.82 (29.57)	–
Days 1 & 2 vs. Other days	–	–	2.60 (7.12)	−1.00 (8.35)	10.41 (14.79)	–
Weight loss attempts						
Day of observation	−.02 (.02)	–	−60.66 (42.23)	−54.80 (43.68)	177.66 (126.40)	−58.38 (48.97)
Weight loss attempts	−.003 (.002)	–	180.53 (117.19)	−159.15 (103.43)	239.60 (385.46)	−73.08 (144.32)
Day of observation × Weight loss attempts	.0004 (.0002) *sr* = .06 *d* = .07 (.02, .13)	–	−29.71 (18.78) *sr* = .13 *d* = −.12 (−.25, .01)	33.13 (20.57) *sr* = .14 *d* = .12 (.00, .24[a])	−54.47 (63.08) *sr* = .16 *d* = −.14 (−.41, .13)	−5.09 (24.99) *sr* = .02 *d* = −.01 (−.13, .11)

TABLE 6 (Continued)

	NHANES *B* (*SE*)	MIDUS *B* (*SE*)	ENACT *B* (*SE*)	IMPACT *B* (*SE*)	FitLink pilot *B* (*SE*)	FitLink full *B* (*SE*)
Day 1 vs. Other days	.0004 (.0002)	–	−29.71 (18.78)	33.13 (20.57)	−54.47 (63.08)	−5.09 (24.99)
Days 1 & 2 vs. Other days	.0002 (.00009)	–	−14.86 (9.39)	16.57 (10.29)	−27.23 (31.54)	−2.54 (12.50)
Social comparison						
Day of observation	–	−4.76 (2.17)	−55.82 (54.35)	–	–	−57.22 (49.28)
Social comparison						
Higher risk (where available)	–	−19.39 (21.56)	69.60 (41.56)	–	–	−45.34 (52.19)
Lower risk (where available)		−2.93 (24.17)				
Day of observation × Social comparison	–	6.49 (3.38) 7.42 (3.77) *sr* = .12 *d* = .14 (.02, .25[a])	−7.93 (6.92) *sr* = .10 *d* = −.12 (−.30, .05)	–	–	6.74 (8.97) *sr* = .07 *d* = −.05 (−.07, .18)
Day 1 vs. Other days	–	−3.42 (1.68)	−7.93 (6.92)	–	–	6.74 (8.97)
Days 1 & 2 vs. Other days	–	−1.71 (.84)	−3.96 (3.46)	–	–	3.37 (4.48)

Note: PA outcome for NHANES: MIMS/min, for MIDUS: activity counts/min; for ENACT, IMPACT, FitLink Pilot, and FitLink Full: steps/day; All models controlled for BMI and day of week (weekday vs. weekend) when available; Standard effect sizes (*sr*) and Cohen's *d* (90% CI) were calculated to test significance.

[a] *sr* ≥ .25 or *d* upper-limit CI ≥ .20.

*$p < .01$; **$p < .0001$.

Aim 3: Design characteristics and PA measurement reactivity

There were also few differences in reactivity patterns between observation-only versus intervention study design. Effect sizes were lower in observational studies (*sr*s = .06–.10; *d*s = −.09 to −.06) relative to clinical trials (*sr*s = .10–.26; *d*s = −.12 to .23; see Tables 5 and 6), but this was primarily driven by one clinical trial with a small sample that showed moderate effects (FitLink Pilot; *sr* = .26; *d* = .23, 90% CI: −.03, .50). Remaining studies had small effect sizes within a narrow range (*sr*s = .10–.12; *d*s = −.12 to .12). There were also no differences in PA patterns across study designs with respect to moderators, with the exception of gender. Effects of gender on measurement reactivity were smaller on average in observational studies (*sr*s = .02 to .09; *d*s = −.08 to .01) relative to intervention studies (*sr*s = .01–.31; *d*s = −.28–.01). This difference was also primarily attributed to the FitLink Pilot, where men *increased* their steps over the course of the observation period (opposite of typical reactivity pattern; *sr* = .31; *d* = −.28, 90% CI: −.55, −.02; see Table 5).

Similarly, there were no meaningful differences in PA patterns with respect to device location or availability of device feedback. Average effect sizes were highly similar for studies using hip-worn (average *sr* = .11) and wrist-worn devices (average *sr* = .13); moderators of PA patterns were also similar between studies using different devices. The only noteworthy difference was for depressive symptoms. The association between depressive symptoms and PA patterns was slightly higher on average among studies using hip-worn (average *sr* = .12, *d* = .11) versus wrist-worn devices (average *sr* = .06, *d* = .02). PA patterns were also similar between studies with devices that provide PA feedback (vs, do not provide feedback), with motivation as an exception. Across the three studies that measured these experiences, the influence of motivation on PA patterns was larger in one study that provided feedback (FitLink Pilot; *sr* = .13, *d* = .11, 90% CI: −.15, .38) compared with two studies that did not provide device feedback (ENACT, IMPACT; *sr*s = .01–.03; *d*s = .01–.03). Overall, however, these differences were modest and thus unlikely to meaningfully bias PA conclusions.

DISCUSSION

PA measurement reactivity is a possible source of bias that can complicate assessment and consequent understanding of ambulatory PA behaviour. This is a particular concern with respect to determining who should be the targets of PA intervention efforts and how well these interventions work to increase PA (Baumann et al., 2018). As noted, concern about PA measurement reactivity has led to recommendations for procedures that can introduce considerable burden for both researchers and participants. Although existing evidence suggests that PA measurement reactivity may be present for some individuals (Arigo & König, 2024; Clemes & Deans, 2012), the extent of its effect on PA estimates remains unclear, as does who is most affected and under what circumstances. To address these questions in a population that commonly receives PA resources and intervention, the goal of the present study was to estimate PA measurement reactivity and potential moderators of its impact on PA behaviour among adults ages 40–60 with CVD risk factors, across six existing datasets. Our coordinated analysis approach enabled examination of reactivity within and across studies using a consistent modelling strategy for rigorous development of conclusions.

Findings from this coordinated analysis, including those of supplemental equivalence tests, show limited evidence of PA measurement reactivity: there was no pattern that indicated reactivity in five of the six studies included. Although a statistically significant decline in PA across days was observed for NHANES, this is attributable to its considerable sample size (*n* = 1208, *k* = 7808 observations), which provided high sensitivity for detecting even very small effects. Importantly, the overall impact of measurement reactivity on estimates of PA was minimal (and likely, inconsequential) across all studies, with the exception of a small pilot trial. Given the sample size, findings from this trial should be interpreted with caution. As noted, an advantage of coordinated analysis is its ability to speak to replicability and generalizability: statistical procedures and inference thresholds are the same across studies that were

conducted for different purposes, and conclusions rest on patterns across or between studies (vs. for individual studies). Thus, although findings from the FitLink pilot trial are noteworthy (see below for further discussion), the broader pattern suggests that they are anomalous and likely due to a small and non-representative sample.

Overall, our findings are consistent with meta-analytic work in this area (König et al., 2022) as well as with empirical evidence from similar populations of adults (Maher et al., 2024), suggesting that the influence of measurement reactivity on PA estimates is negligible. Of note, prior work seems to focus exclusively on initial elevation (i.e., decreases over time), rather than other patterns that could indicate a response to measurement; in the present study, we observed several instances of later elevation (i.e., increases over time), which warrant additional attention. Prior research has also highlighted the role of individual differences such as social desirability, self-monitoring tendencies, or concern with others' perceptions in shaping reactivity to self-monitoring or behaviour in observational settings (Ambwani & Chmielewski, 2013). In the present study, we observed little evidence that subsets of participants are particularly prone to reactivity. Complex interactions indicated gender differences in the association between depression and PA patterns and between social comparison and PA patterns. Yet, these effects were small and were only seen in one study each (of six that measured depressive symptoms and three that measured social comparison). Consequently, there is little evidence that these represent important and widespread individual differences in the likelihood of reactivity responses.

Similarly, evidence from the present study indicates that the effects of study design and PA measurement device are minimal. There were no consistent, minimally meaningful differences in PA patterns between observational and intervention studies or those using different types of devices. The differences observed herein were driven by one intervention study with a moderate effect size (FitLink Pilot). This study also had a very small sample size, as noted, and had very few men enrolled as participants. Further, reactivity may have been greater in this study due to fewer required pre-intervention study activities (e.g., orientation sessions) before the PA measurement period, relative to other intervention studies (ENACT, FitLink Full). PA behaviour and measurement may have been more salient for these participants compared with those who communicated with the research team for a longer period (and thus, had more time to get used to researcher observation; Motl et al., 2012). Similarly, there was little variation in reactivity effects with respect to device differences. Exceptions to this were greater influence of depressive symptoms in studies using hip- versus wrist-worn devices and greater influence of motivation in a study providing device feedback. However, as differences in effect sizes were small, there is little evidence of meaningful effects of reactivity on PA estimates.

If PA measurement reactivity were ubiquitous and affecting PA estimates across PA research studies, as has been suggested (cf. French et al., 2021), it should appear in any study (or at least most) that involves PA measurement across days of observation. Yet, taken together, findings from the present study indicate that measurement reactivity is *not* likely to meaningfully bias PA estimates among midlife adults with CVD risk. Consequently, there is little justification for altering research design or procedures with this population, such as adding initial days of observation that will be excluded from statistical analyses (cf. Clemes & Deans, 2012). Instead, we reiterate encouragement for researchers to statistically test for reactivity patterns in any study that uses device-based PA measurement and to control for study day in analyses if there are small differences across or between days (cf. Arigo & König, 2024; Maher et al., 2024).

We also recommend explicitly introducing the concept of reactivity during participant orientations to studies that use device-based PA measurement, to ensure that they are aware of the potential for changes in their behaviour due to enrolling in the study and knowing that their PA is being measured. For instance, in clinical trials, the overarching objective may be to encourage PA behaviour change, though this change is expected to occur only after a baseline measurement period. It is important to explain the purpose of this initial assessment and the potential for reactivity, as participants often experience peak motivation for behaviour change at the start of a programme (and thus may engage in PA at levels they cannot sustain; Neve et al., 2010). In observational studies, greater emphasis on continuing typical PA during the measurement period to capture accurate behaviour in the natural environment

could be useful. Many participants have a general understanding that PA is beneficial for health and may use their participation as an opportunity to try to meet recommended guidelines, as someone else is watching (McCambridge et al., 2014). Reassurance that this is not expected, or explicit discouragement to change behaviour simply because they are enrolled in a study, may effectively mitigate any potential, minimal effect of reactivity.

Strengths, limitations, and future directions

A major strength of the current study is our use of pre-specified, coordinated multilevel analysis across six different studies, which we supplemented with equivalence testing. This allowed us to represent a wide range of participant characteristics and PA measurement approaches from nearly two decades of research (2005–2023), increasing the generalizability and robustness of the current findings. The use of existing data also enabled us to test the phenomenon using multiple PA metrics (e.g., steps per day, MIMS units, activity counts). This approach is responsive to calls for examining reactivity in PA outcomes other than steps per day, to gain deeper insights into the potential for reactivity to bias estimates of PA behaviour (French & Sutton, 2010; König et al., 2022). Notably, however, PA outcomes varied across studies; MIMS units and total activity counts cannot be converted to estimates of daily steps or vice versa, to make direct comparisons across all studies (Baga et al., 2025; Lee et al., 2023). We addressed this directly using standardized effect sizes with equivalence tests based on confidence intervals; however, comparison in the original unit across studies would be desirable, as this would also speak directly to the potential for clinical or practical significance of potential reactivity patterns. As considerable intervention research for midlife adults with elevated CVD risk focuses on increasing structured exercise or reducing sedentary time (Arigo et al., 2022; Mosalman Haghighi et al., 2018), examining reactivity in parameters of PA behaviour other than overall movement in this group is also important, though it is unlikely to yield meaningfully different results (see Arigo & König, 2024; Maher et al., 2024).

In addition, it is unclear whether participants all used study-assigned devices or their own device (specifically in FitLink Pilot and FitLink Full), or whether participants had prior experience using a commercially available PA device (i.e., Fitbit, pedometer). König et al. (2025) recently tested the role of researcher observation on PA patterns, by randomly assigning participants who already used their own PA monitoring device to a high-salience or a low-salience observation condition. There were no meaningful differences between pre- and post-enrollment steps per day, or between conditions, suggesting that reactivity to PA observation is minimal for adults who already use a PA monitor (König et al., 2025). In the general population of US adults, use of personal PA monitoring devices increased during the period of data collection across studies (i.e., 2005–2023); however, other studies were published during this period that did show evidence of PA measurement reactivity (e.g., Baumann et al., 2018; Clemes & Deans, 2012; Motl et al., 2012; meta-analysis by König et al., 2022), and users consistently tend to be healthier and more physically active than non-users (Friel & Garber, 2020). Consequently, changes in personal PA device use are unlikely to explain findings for the specific population in question (i.e., adults in midlife with risk factors for CVD, whose PA is limited), and we saw no evidence of change in reactivity patterns over time. In future studies, researchers may consider participants' prior experiences with self-monitoring devices, as exposure to these devices outside of a research study could influence their response to the study-issued device and potentially influence PA measurement reactivity in comparison to those that had no prior exposure. The introduction of a new device could play a larger role in PA measurement reactivity when compared with observation alone (König et al., 2025).

Conclusion

PA measurement reactivity has been a concern for scientists and practitioners studying PA promotion. Fortunately, findings from the present study indicate that these effects are unlikely to confound PA

estimates or require intensive changes to measurement procedures or interpretation of results with adults in midlife who have CVD risk factors. Controlling for observation day in analyses (as appropriate) and encouraging participants not to change behaviour during measurement periods should suffice for addressing any unique reactivity patterns. Continued investigation of PA reactivity, particularly for those who may be more reactive to measurement, will yield further nuance.

AUTHOR CONTRIBUTIONS

Danielle Arigo: Conceptualization; funding acquisition; writing – original draft; methodology; writing – review and editing; project administration; supervision; formal analysis. **Kiri Baga:** Writing – original draft; validation; data curation; writing – review and editing; investigation. **Amanda L. Folk:** Writing – original draft; writing – review and editing; formal analysis; visualization. **Gabrielle M. Salvatore:** Writing – original draft; writing – review and editing. **Iris Bercovitz:** Writing – original draft; writing – review and editing. **Ria Singh:** Writing – original draft; writing – review and editing. **Laura M. König:** Writing – original draft; writing – review and editing; methodology. **Meghan L. Butryn:** Conceptualization; writing – review and editing; resources; investigation. **Jacqueline A. Mogle:** Conceptualization; methodology; formal analysis; data curation.

ACKNOWLEDGEMENTS

The authors would like to thank Raj Harsora and Natasha DeMeo, M.Sc. for their assistance with data acquisition and management, and Joann Kandavalli for her assistance with manuscript preparation.

FUNDING INFORMATION

This work was supported by the US National Institutes of Health under NHLBI R03160602 and a competitive administrative supplement from the Office of Research on Women's Health (PI: D. Arigo).

CONFLICT OF INTEREST STATEMENT

The authors confirm that they have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

Data from NHANES and MIDUS are publicly available from the Inter-university Consortium for Political and Social Research (https://www.icpsr.umich.edu/web/pages). Data from the clinical trials described above are available upon reasonable request to Meghan L. Butryn, PhD (mlb34@drexel.edu).

ORCID

Danielle Arigo https://orcid.org/0000-0002-7807-5913
Ria Singh https://orcid.org/0009-0001-7827-1041
Laura M. König https://orcid.org/0000-0003-3655-8842

REFERENCES

Aguiar, E. J., Turner, D. T., Pleuss, J. D., Zheng, P., Benitez, C. J., & Ducharme, S. W. (2024). Daily and peak monitor independent movement summary (MIMS) values associated with metabolic syndrome: NHANES 2011-12 and 2013-14. *Scandinavian Journal of Medicine & Science in Sports*, *34*(11), e14762. https://doi.org/10.1111/sms.14762

Ambwani, S., & Chmielewski, J. F. (2013). Weighing the evidence: Social desirability, eating disorder symptomatology, and accuracy of self-reported body weight among men and women. *Sex Roles*, *68*, 474–483. https://doi.org/10.1007/s11199-012-0244-1

Arigo, D., & König, L. M. (2024). Examining reactivity to the measurement of physical activity and sedentary behavior among women in midlife with elevated risk for cardiovascular disease. *Psychology & Health*, *39*(3), 319–335. https://doi.org/10.1080/08870446.2022.2055024

Arigo, D., Romano, K. A., Pasko, K., Travers, L., Ainsworth, M. C., Jackson, D. A., & Brown, M. M. (2022). A scoping review of behavior change techniques used to promote physical activity among women in midlife. *Frontiers in Psychology*, *13*, 855749. https://doi.org/10.3389/fpsyg.2022.855749

Baga, K., Salvatore, G. M., Bercovitz, I., Folk, A. L., Singh, R., König, L. M., Butryn, M. L., Mogle, J. A., & Arigo, D. (2025). Physical activity measurement reactivity among midlife adults with elevated risk for cardiovascular disease: Protocol for coordinated analyses across six studies. *JMIR Research Protocols*, *14*, e67438. https://doi.org/10.2196/67438

Barta, W. D., Tennen, H., & Litt, M. D. (2012). Measurement reactivity in diary research. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 108–123). The Guilford Press.

Baumann, S., Groß, S., Voigt, L., Ullrich, A., Weymar, F., Schwaneberg, T., Dörr, M., Meyer, C., John, U., & Ulbricht, S. (2018). Pitfalls in accelerometer-based measurement of physical activity: The presence of reactivity in an adult population. *Scandinavian Journal of Medicine & Science in Sports*, *28*(3), 1056–1063. https://doi.org/10.1111/sms.12977

Beck, A. T., Steer, R. A., & Brown, G. (1996). Beck depression inventory–II (BDI-II) [database record]. *APA PsycTests*. https://doi.org/10.1037/t00742-000

Bell, B. M., Alam, R., Alshurafa, N., Thomaz, E., Mondol, A. S., de la Haye, K., Stankovic, J. A., Lach, J., & Spruijt-Metz, D. (2020). Automatic, wearable-based, in-field eating detection approaches for public health research: A scoping review. *Npj Digital Medicine*, *3*, 38. https://doi.org/10.1038/s41746-020-0246-2

Brim, O. G., Ryff, C. D., & Kessler, R. C. (2019). *How healthy are we?: A National Study of well-being at midlife*. University of Chicago Press.

Buunk, B. P., & Ybema, J. F. (1997). Social comparisons and occupational stress: The identification-contrast model. In B. P. Buunk & F. X. Gibbons (Eds.), *Health, coping, and well-being: Perspectives from social comparison theory* (pp. 359–388). Lawrence Erlbaum Associates Publishers.

Centers for Disease Control. (2024). NHANES questionnaires, datasets, and related documentation. https://wwwn.cdc.gov/nchs/nhanes/

Chekroud, S. R., Gueorguieva, R., Zheutlin, A. B., Paulus, M., Krumholz, H. M., Krystal, J. H., & Chekroud, A. M. (2018). Association between physical exercise and mental health in 1·2 million individuals in the USA between 2011 and 2015: A cross-sectional study. *The Lancet Psychiatry*, *5*(9), 739–746. https://doi.org/10.1016/S2215-0366(18)30227-X

Christiansen, L. B., Koch, S., Bauman, A., Toftager, M., Bjørk Petersen, C., & Schipperijn, J. (2023). Device-based physical activity measures for population surveillance—Issues of selection bias and reactivity. *Frontiers in Sports and Active Living*, *5*, 1236870. https://doi.org/10.3389/fspor.2023.1236870

Clemes, S. A., & Deans, N. K. (2012). Presence and duration of reactivity to pedometers in adults. *Medicine and Science in Sports and Exercise*, *44*(6), 1097–1107. https://doi.org/10.1249/MSS.0b013e318242a377

Cohen, B. E., Edmondson, D., & Kronish, I. M. (2015). State of the art review: Depression, stress, anxiety, and cardiovascular disease. *American Journal of Hypertension*, *28*(11), 1295–1302. https://doi.org/10.1093/ajh/hpv047

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Cooke, R., & Jones, A. (2017). Recruiting adult participants to physical activity intervention studies using sport: A systematic review. *BMJ Open Sport & Exercise Medicine*, *3*(1), e000231. https://doi.org/10.1136/bmjsem-2017-000231

Davis, R. E., & Loprinzi, P. D. (2016). Examination of accelerometer reactivity among a population sample of children, adolescents, and adults. *Journal of Physical Activity & Health*, *13*(12), 1325–1332. https://doi.org/10.1123/jpah.2015-0703

Edney, S., Ryan, J. C., Olds, T., Monroe, C., Fraysse, F., Vandelanotte, C., Plotnikoff, R., Curtis, R., & Maher, C. (2019). User engagement and attrition in an app-based physical activity intervention: Secondary analysis of a randomized controlled trial. *Journal of Medical Internet Research*, *21*(11), e14645. https://doi.org/10.2196/14645

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532–538. https://doi.org/10.1037/a0015808

Foote, S. J., Wadsworth, D. D., Brock, S., Hastie, P., & Cooper, C. K. (2017). The effect of a wrist worn accelerometer on children's in-school and out-of-school physical activity levels. *Swedish Journal of Scientific Research*, *33*, 1–6.

French, D. P., & Sutton, S. (2010). Reactivity of measurement in health psychology: How much of a problem is it? What can be done about it? *British Journal of Health Psychology*, *15*(3), 453–468. https://doi.org/10.1348/135910710X492341

French, D. P., Miles, L. M., Elbourne, D., Farmer, A., Gulliford, M., Locock, L., Sutton, S., McCambridge, J., & MERIT Collaborative Group. (2021). Reducing bias in trials due to reactions to measurement: Experts produced recommendations informed by evidence. *Journal of Clinical Epidemiology*, *139*, 130–139. https://doi.org/10.1016/j.jclinepi.2021.06.028

Friel, C. P., & Garber, C. E. (2020). Who uses wearable activity trackers and why? A comparison of former and current users in the United States. *American Journal of Health Promotion*, *34*(7), 762–769. https://doi.org/10.1177/0890117120919366

Gibbons, F. X., & Buunk, B. P. (1999). Individual differences in social comparison: Development of a scale of social comparison orientation. *Journal of Personality and Social Psychology*, *76*(1), 129–142. https://doi.org/10.1037/0022-3514.76.1.129

Graham, E. K., Willroth, E. C., Weston, S. J., Muniz-Terrera, G., Clouston, S. A. P., Hofer, S. M., Mroczek, D. K., & Piccinin, A. M. (2022). Coordinated data analysis: Knowledge accumulation in lifespan developmental psychology. *Psychology and Aging*, *37*(1), 125–135. https://doi.org/10.1037/pag0000612

Hansen, B. H., Holme, I., Anderssen, S. A., & Kolle, E. (2013). Patterns of objectively measured physical activity in normal weight, overweight, and obese individuals (20–85 years): A cross-sectional study. *PLoS One*, *8*(1), e53044. https://doi.org/10.1371/journal.pone.0053044

Haynes, A., & Robinson, E. (2019). Who are we testing? Self-selection bias in laboratory-based eating behaviour studies. *Appetite*, *141*, 104330. https://doi.org/10.1016/j.appet.2019.104330

Hilden, P., Schwartz, J. E., Pascual, C., Diaz, K. M., & Goldsmith, J. (2023). How many days are needed? Measurement reliability of wearable device data to assess physical activity. *PLoS One*, *18*(2), e0282162. https://doi.org/10.1371/journal.pone.0282162

Hilgenkamp, T., Van Wijck, R., & Evenhuis, H. (2012). Measuring physical activity with pedometers in older adults with intellectual disability: Reactivity and number of days. *Intellectual and Developmental Disabilities*, *50*(4), 343–351. https://doi.org/10.1352/1934-9556-50.4.343

Hill, N. L., Bhargava, S., Bratlee-Whitaker, E., Turner, J. R., Brown, M. J., & Mogle, J. (2021). Longitudinal relationships between subjective cognitive decline and objective memory: Depressive symptoms mediate between-person associations. *Journal of Alzheimer's Disease*, *83*(4), 1623–1636. https://doi.org/10.3233/JAD-210230

Hofer, S. M., & Piccinin, A. M. (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods*, *14*(2), 150–164. https://doi.org/10.1037/a0015566

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change* (1st ed.). Routledge. https://doi.org/10.4324/9781315744094

Hox, J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge. https://doi.org/10.4324/9781315650982

John, D., Tang, Q., Albinali, F., & Intille, S. (2019). An open-source monitor-independent movement summary for accelerometer data processing. *Journal for the Measurement of Physical Behaviour*, *2*(4), 268–281. https://doi.org/10.1123/jmpb.2018-0068

Karvinen, S., Jergenson, M. J., Hyvärinen, M., Aukee, P., Tammelin, T., Sipilä, S., Kovanen, V., Kujala, U. M., & Laakkonen, E. K. (2019). Menopausal status and physical activity are independently associated with cardiovascular risk factors of healthy middle-aged women: Cross-sectional and longitudinal evidence. *Frontiers in Endocrinology*, *10*, 589. https://doi.org/10.3389/fendo.2019.00589

Klenk, J., Peter, R. S., Rapp, K., Dallmeier, D., Rothenbacher, D., Denkinger, M., Büchele, G., Becker, T., Böhm, B., Scharffetter-Kochanek, K., Stingl, J., Koenig, W., Riepe, M., Peter, R., Geiger, H., Ludolph, A., von Arnim, C., Nagel, G., Weinmayr, G., & Laszlo, R. (2019). Lazy Sundays: Role of day of the week and reactivity on objectively measured physical activity in older people. *European Review of Aging and Physical Activity*, *16*(1), 18. https://doi.org/10.1186/s11556-019-0226-1

König, L. M., Allmeta, A., Christlein, N., Van Emmenis, M., & Sutton, S. (2022). A systematic review and meta-analysis of studies of reactivity to digital in-the-moment measurement of health behaviour. *Health Psychology Review*, *16*(4), 551–575. https://doi.org/10.1080/17437199.2022.2047096

König, L. M., Pasko, K., Baga, K., Harsora, R., & Arigo, D. (2025). Isolating the role of researcher observation on reactivity to the measurement of physical activity. *Applied Psychology. Health and Well-Being*, *17*(1), e12630. https://doi.org/10.1111/aphw.12630

Kraus, W. E., Powell, K. E., Haskell, W. L., Janz, K. F., Campbell, W. W., Jakicic, J. M., Troiano, R. P., Sprow, K., Torres, A., Piercy, K. L., & Physical Activity Guidelines Advisory Committee. (2019). Physical activity, all-cause and cardiovascular mortality, and cardiovascular disease. *Medicine and Science in Sports and Exercise*, *51*(6), 1270–1281. https://doi.org/10.1249/MSS.0000000000001939

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, *32*(9), 509–515. https://doi.org/10.3928/0048-5713-20020901-06

Lee, I.-M., Moore, C. C., & Evenson, K. R. (2023). Maximizing the utility and comparability of accelerometer data from large-scale epidemiologic studies. *Journal for the Measurement of Physical Behaviour*, *6*(1), 6–12. https://doi.org/10.1123/jmpb.2022-0035

Levesque, C. S., Williams, G. C., Elliot, D., Pickering, M. A., Bodenhamer, B., & Finley, P. J. (2007). Validating the theoretical structure of the treatment self-regulation questionnaire (TSRQ) across three different health behaviors. *Health Education Research*, *22*(5), 691–702. https://doi.org/10.1093/her/cyl148

Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education*, *6*(1), 8. https://doi.org/10.1186/s40536-018-0061-2

Maher, J. P., Arigo, D., Baga, K., Salvatore, G. M., Pasko, K., Hudgins, B. L., & König, L. M. (2024). Measurement reactivity in ecological momentary assessment studies of movement-related behaviors. *Journal for the Measurement of Physical Behaviour*, *7*(1), jmpb.2023-0035.

Matthews, K. A., Crawford, S. L., Chae, C. U., Everson-Rose, S. A., Sowers, M. F., Sternfeld, B., & Sutton-Tyrrell, K. (2009). Are changes in cardiovascular disease risk factors in midlife women due to chronological aging or to the menopausal transition? *Journal of the American College of Cardiology*, *54*(25), 2366–2373. https://doi.org/10.1016/j.jacc.2009.10.009

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology*, *67*(3), 267–277.

Mogle, J., Hill, N. L., & Turner, J. R. (2021). Individual differences and features of self-reported memory lapses as risk factors for alzheimer disease among adults aged 50 years and older: Protocol for a coordinated analysis across two longitudinal data sets. *JMIR Research Protocols*, *10*(5), e25233. https://doi.org/10.2196/25233

Mosalman Haghighi, M., Mavros, Y., & Fiatarone Singh, M. A. (2018). The effects of structured exercise or lifestyle behavior interventions on long-term physical activity level and health outcomes in individuals with type 2 diabetes: A systematic review, meta-analysis, and meta-regression. *Journal of Physical Activity and Health*, *15*(9), 697–707. https://doi.org/10.1123/jpah.2017-0589

Motl, R. W., McAuley, E., & Dlugonski, D. (2012). Reactivity in baseline accelerometer data from a physical activity behavioral intervention. *Health Psychology*, *31*(2), 172–175. https://doi.org/10.1037/a0025965

Neve, M. J., Collins, C. E., & Morgan, P. J. (2010). Dropout, nonusage attrition, and pretreatment predictors of nonusage attrition in a commercial web-based weight loss program. *Journal of Medical Internet Research*, *12*(4), e69. https://doi.org/10.2196/jmir.1640

Ozdoba, R., Corbin, C., & Le Masurier, G. (2004). Does reactivity exist in children when measuring activity levels with unsealed pedometers? *Pediatric Exercise Science*, *16*(2), 158–166. https://doi.org/10.1123/pes.16.2.158

Panjeh, S., Nordahl-Hansen, A., & Cogo-Moreira, H. (2023). Establishing new cutoffs for Cohen's d: An application using known effect sizes from trials for improving sleep quality on composite mental health. *International Journal of Methods in Psychiatric Research*, *32*(3), e1969. https://doi.org/10.1002/mpr.1969

Poulton, A., Pan, J., Bruns, L. R., Jr., Sinnott, R. O., & Hester, R. (2019). A smartphone app to assess alcohol consumption behavior: Development, compliance, and reactivity. *JMIR mHealth and uHealth*, *7*(3), e11157. https://doi.org/10.2196/11157

Prewitt, S. L., Hannon, J. C., & Brusseau, T. A. (2013). Children and pedometers: A study in reactivity and knowledge. *International Journal of Exercise Science*, *6*(3), 230–235. https://doi.org/10.70252/AUIG2070

Reeves, W. C., Strine, T. W., Pratt, L. A., Thompson, W., Ahluwalia, I., Dhingra, S. S., McKnight-Eily, L. R., Harrison, L., D'Angelo, D. V., Williams, L., Morrow, B., Gould, D., Safran, M. A., & Centers for Disease Control and Prevention (CDC). (2011). Mental illness surveillance among adults in the United States. *MMWR Supplements*, *60*(3), 1–29.

Rodgers, J. L., Jones, J., Bolleddu, S. I., Vanthenapalli, S., Rodgers, L. E., Shah, K., Karia, K., & Panguluri, S. K. (2019). Cardiovascular risks associated with gender and aging. *Journal of Cardiovascular Development and Disease*, *6*(2), 19. https://doi.org/10.3390/jcdd6020019

Rutledge, T., & Loh, C. (2004). Effect sizes and statistical testing in the determination of clinical significance in behavioral medicine research. *Annals of Behavioral Medicine*, *27*(2), 138–145. https://doi.org/10.1207/s15324796abm2702_9

Ryff, C. D., Weinstein, M., & Seeman, T. E. (2010). National Survey of Midlife Development in the United States (MIDUS II): Biomarker Project, 2004–2009. https://doi.org/10.3886/ICPSR29282.v10

Sassarini, J. (2016). Depression in midlife women. *Maturitas*, *94*, 149–154. https://doi.org/10.1016/j.maturitas.2016.09.004

Simon, G. E., Ludman, E. J., Linde, J. A., Operskalski, B. H., Ichikawa, L., Rohde, P., Finch, E. A., & Jeffery, R. W. (2008). Association between obesity and depression in middle-aged women. *General Hospital Psychiatry*, *30*(1), 32–39. https://doi.org/10.1016/j.genhosppsych.2007.09.001

Tang, J. S., Haslam, R. L., Ashton, L. M., Fenton, S., & Collins, C. E. (2022). Gender differences in social desirability and approval biases, and associations with diet quality in young adults. *Appetite*, *175*, 106035. https://doi.org/10.1016/j.appet.2022.106035

Tudor-Locke, C., Brashear, M. M., Johnson, W. D., & Katzmarzyk, P. T. (2010). Accelerometer profiles of physical activity and inactivity in normal weight, overweight, and obese U.S. men and women. *International Journal of Behavioral Nutrition and Physical Activity*, *7*, 60. https://doi.org/10.1186/1479-5868-7-60

Vincent, S. D., & Pangrazi, R. P. (2002). Does reactivity exist in children when measuring activity levels with pedometers? *Pediatric Exercise Science*, *14*(1), 56–63. https://doi.org/10.1123/pes.14.1.56

Wadden, T. A., & Foster, G. D. (2006). Weight and Lifestyle Inventory (WALI). *Obesity*, *14*(S3), 99S–118S. https://doi.org/10.1038/oby.2006.289

Waters, L. A., Galichet, B., Owen, N., & Eakin, E. (2011). Who participates in physical activity intervention trials? *Journal of Physical Activity and Health*, *8*(1), 85–103. https://doi.org/10.1123/jpah.8.1.85

Zhu, X., & Haegele, J. A. (2019). Reactivity to accelerometer measurement of children with visual impairments and their family members. *Adapted Physical Activity Quarterly: APAQ*, *36*(4), 492–500. https://doi.org/10.1123/apaq.2019-0040

How to cite this article: Arigo, D., Baga, K., Folk, A. L., Salvatore, G. M., Bercovitz, I., Singh, R., König, L. M., Butryn, M. L., & Mogle, J. A. (2026). Do adults with cardiovascular disease risk show meaningful reactivity to physical activity measurement? Coordinated analysis across six studies. *British Journal of Health Psychology*, *31*, e70063. https://doi.org/10.1111/bjhp.70063