A systematic review and meta-analysis of studies of reactivity to digital in-the-moment measurement of health behaviour

Laura M König[1,2*], Anila Allmeta[1], Nora Christlein[1], Miranda Van Emmenis[2], & Stephen Sutton[2]

[1] Faculty of Life Sciences: Food, Nutrition and Health, University of Bayreuth

[2] Behavioural Science Group, Primary Care Unit, Department of Public Health and Primary Care, University of Cambridge

* Corresponding author

Dr. Laura M. König
University of Bayreuth
Faculty 7/ Campus Kulmbach
Fritz-Hornschuch-Straße 13
93526 Kulmbach
Email: laura.koenig@uni-bayreuth.de

Author contributions

SS and LK conceived of the study and developed the research question. LK developed the search strategy with input from SS, AA and MVE. Searches were conducted by LK; screening was conducted by LK, AA, MVE and NC; data was extracted by LK and NC; study quality was appraised by AA, LK and MVE. Data was analysed and interpreted, and the manuscript was written by LK with input from all authors. All authors approved the final version of the manuscript.

Competing interest

The authors declare no competing interests.

Data availability statement

A list of all full texts screened, the raw data extracted from the included publications, effect sizes and Jamovi analysis files can be obtained from https://osf.io/qnmvj/. The search strategy, a list of all extracted information, effect sizes for the experimental studies and results of the risk of bias assessment can be found in the online supplement.

Abstract

Self-report measures of health behaviour have several limitations including measurement reactivity, i.e. changes in people's behaviour, cognitions or emotions due to taking part in research. This systematic review investigates whether digital in-the-moment measures induce reactivity to a similar extent and why it occurs. Four databases were searched in December 2020. All observational or experimental studies investigating reactivity to digital in-the-moment measurement of a range of health behaviours were included if they were published in English in 2008 or later. Of the 11,723 records initially screened, 30 publications reporting on 31 studies were included in the qualitative synthesis/ 7 studies in the quantitative synthesis. Eighty-one percent of studies focused on reactivity to the measurement of physical activity indicators; small but meaningful pooled effects were found (Cohen's ds: 0.27 to 0.30). Only a small number of studies included other behaviours, yielding mixed results. Digital in-the-moment measurement of behaviour thus may be as prone to reactivity as self-reports in questionnaires. Measurement reactivity may be amplified by (1) ease of changing the behaviour, (2) awareness of being measured and social desirability, and (3) resolving discrepancies between actual and desired behaviour through self-regulation.

Keywords: measurement reactivity, assessment reactivity, physical activity, alcohol consumption, soft drink consumption, smoking

Introduction

In behavioural science, including psychology, self-report measures of behaviour are ubiquitous (1). Self-report measures provide researchers with information on people's behaviour at a comparably low cost, e.g. when data is collected via (online) questionnaires. However, they suffer from several shortcomings. Self-report assessments are usually retrospective one-time measures that require participants to average across many different occasions, inducing a recall bias (2). For instance, Food Frequency Questionnaires typically ask participants to indicate the frequency of consuming certain categories of food within the past months or year (e.g., (3); see also Thompson and Subar (4) for a discussion). Furthermore, responses may be biased because of social desirability, i.e. the participant answering in a way that they feel will satisfy the researcher instead of reporting their actual behaviour or emotions (5). Finally, self-report measures such as questionnaires have been shown to be prone to measurement reactivity (6, 7), i.e. changes in people's behaviour, emotions or cognitions due to being measured as part of a research project (8). Since measurement reactivity has mainly been studied in the context of questionnaire-based studies, it is also referred to as the question-behaviour effect (9) or mere-measurement effect (10, 11). Several recent systematic reviews and meta-analyses have summarised the evidence on the impact of completing a (pen-and-paper or online) questionnaire on a range of behaviours. They consistently report small but significant effect sizes of Cohen's d ranging from 0.06 to 0.28 (12-17). It is thus questioned whether self-report measurements allow researchers to draw sufficiently valid conclusions about human behaviour as well as its determinants and consequences (8).

To overcome these limitations, it is often recommended to use objective measures of behaviour or to reduce the time span between the behaviour occurring and it being assessed by using Ecological Momentary Assessment (18). Because of recent technological developments and reduced costs, the use of digital measurement devices is becoming increasingly popular especially in health behaviour research including health psychology (19). For instance, health-related behaviours such as physical activity and sedentary behaviour are increasingly tracked using wearable devices (20), while consumption behaviours such as eating are increasingly studied using smartphones (21) and body-worn sensors (22). These methods provide detailed insights into health behaviours in daily life and allow individuals to measure behaviour when it occurs, thereby reducing recall bias and increasing the validity of the collected data (18). Accordingly, measuring behaviour immediately when it occurs in daily life using digital devices is assumed to have fewer methodological shortcomings than self-reports assessed via questionnaires. It is unclear, however, whether the validity of behavioural data that is recorded with digital devices is also influenced by research participation effects such as measurement reactivity (8). For example, when using digital measurement devices, people are typically aware of the study context. Accordingly, it could be hypothesised that also digital measurement of behaviour suffers from measurement reactivity.

Measurement reactivity may be especially challenging in the behavioural and health sciences, where digital assessments of behaviour are increasingly used to study determinants of behaviour and to evaluate the effectiveness of interventions (e.g., König et al. (21), Degroote et al. (23)). In contrast to self-monitoring, which is used as a Behaviour Change Technique to deliberately induce changes in behaviour through recording (24), measurement reactivity is usually undesired since it distorts the study findings. For example, recording a behaviour might induce reflecting on the behaviour, which might increase the likelihood of behaviour change independent of intervention components (7, 25). This, in turn, may lead to ineffective interventions being implemented on a larger scale, so creating unnecessary costs and preventing more effective interventions from being implemented. On the other hand, assessment tools might introduce systematic bias that conceals true intervention effects, e.g. when a self-monitoring device used for the intervention is also used for baseline measurements in intervention and control groups. The tool may impact behaviour in all participants and thus lead to the erroneous conclusion that the intervention was unsuccessful (26) (see also (27) for a discussion). Taking potential measurement reactivity into account when analysing behavioural data by introducing it as a model parameter, may be crucial; however, this is only possible if an estimate of the effect is available (27, 28). The present systematic review updates an earlier rapid review (see (27, 29) for summaries) to synthesise the evidence on and the magnitude of reactivity to digital in-the-moment measurement of health behaviour, to guide future research activities on measurement reactivity and extend existing guidance on reducing measurement reactivity in behavioural and clinical research (27).

Methods

A protocol was developed following the PRISMA 2009 guidelines (30) and registered on PROSPERO (registration number CRD42021221933) prior to conducting this systematic review. This report follows the updated PRISMA 2020 guidelines (31). Raw data and analysis scripts are available on the project's Open Science Framework page: https://osf.io/qnmvj/.

Inclusion and exclusion criteria

Any observational or experimental study using a digital (e.g., smartphone app, pedometer, medication event monitoring system) tool to assess behavioural data repeatedly in daily life was included. Study protocols, systematic reviews and meta-analyses were excluded. Studies were also excluded if assessments were not digital (e.g., paper diaries). Behaviours were restricted to the following health behaviours which are among the leading health risk factors (32): alcohol consumption, dental care, diet, medication adherence, physical activity, sedentary behaviour, smoking. Studies were included if they aimed to investigate reactivity to the measurement, i.e. changes in behaviour due to being measured as part of a research project (8), comparing records either between different conditions (e.g., comparing different devices) or within participants across conditions or over time. Self-monitoring interventions or interventions providing feedback that aimed to change participants' behaviour through tracking were excluded. There were no restrictions regarding the participants' age or health status. Studies had to be published in peer-reviewed journals and written in English; accordingly, studies published in any other language as well as theses and preprints were excluded.

Search strategy

An electronic search strategy was developed based on the inclusion criteria. It included keywords related to the different health behaviours and tools for their assessment (e.g., physical activity, pedomet*) and measurement/ assessment reactivity. The strategy was initially developed for Pubmed (incl. MEDLINE) and adapted for PsycINFO, Embase and Web of Science Core Collection (see Appendix A for the strategies developed for the different databases). All databases were searched from 2008 to 1st December 2020, when the search was conducted. The publication date was restricted to 2008 onwards since research in digital assessment of health behaviour has accelerated since then through the development of smartphones, and the vast majority of papers in this field have been published in the 2010s (19). Reference lists of all eligible papers were hand searched and forward citation tracking using Google Scholar was conducted to identify further eligible publications. Moreover, emails were sent to the members' mailing lists of the European Health Psychology Society, the German Psychological Society and the British Psychological Society to identify further eligible work.

Study selection

All records retrieved from the database searches were imported into Covidence systematic review software (Veritas Health Innovation, Melbourne, Australia; available at www.covidence.org). Duplicates were removed before titles and abstracts were screened independently by two authors, coding articles as provisionally eligible or excluded according to the inclusion and exclusion criteria. Disagreements were resolved by discussion. Afterwards, full texts were screened independently by two authors and coded as eligible or excluded. Again, disagreements were resolved by discussion. A PRISMA flow diagram (31) documents the flow of records (see Figure 1).

Data extraction and synthesis

For all eligible studies, two authors independently extracted study characteristics relevant to categorising and describing the studies (e.g., target behaviour[s], study design, description of the digital assessment tool[s], description of condition[s], moderators) and outcomes (effect size[s] or relevant indices needed for calculating the effect size, overall conclusion regarding the presence of measurement reactivity) in a form that was developed based on a previous rapid review ((29); see Appendix B for the full list of information extracted). Discrepancies were identified and resolved by discussion. Study authors were contacted to obtain key unpublished outcome data. Data on all studies were synthesised narratively.

In addition, a meta-analysis was conducted for experimental studies on reactivity to measuring all physical activity indicators combined and for steps only using MAJOR – Meta-Analysis for Jamovi Version 1.2.0 (33) with Jamovi 1.6.23 (34). Random effects models were computed to calculate pooled effect sizes. Since all experimental studies compared continuous data collected from at least two groups, Cohen's *d* is reported. If available, means and standard deviations were extracted from the studies to calculate Cohen's *d* (35) following the

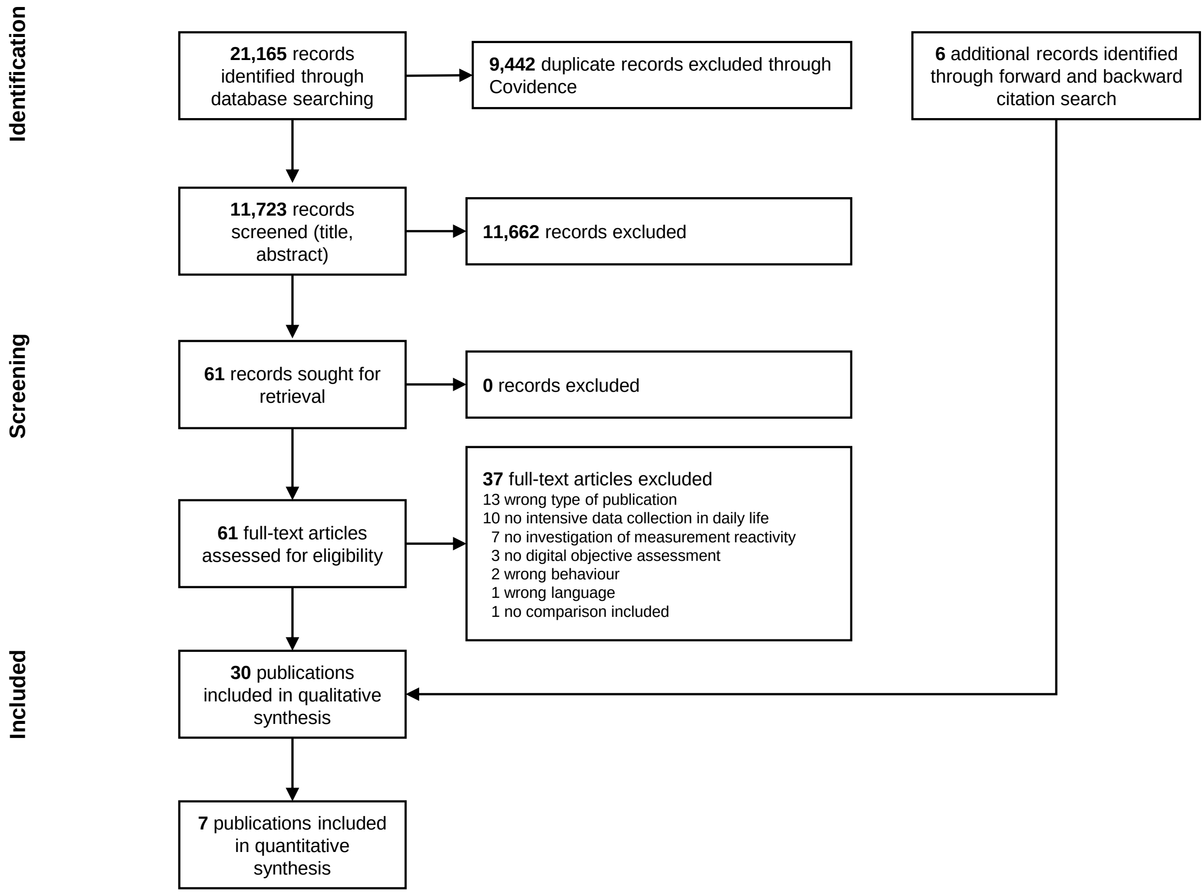

Figure 1. PRISMA flow chart.

recommendations outlined in Borenstein et al. (36) and The Cochrane Collaboration (37). If the information provided in the publication was not sufficient to calculate Cohen's *d*, authors were contacted and asked to provide the effect size or raw data. For three studies, no information about effect sizes could be obtained (38-40). Heterogeneity of effect sizes was evaluated using I^2 as recommended by Higgins et al. (41). Due to a small number of studies for individual combinations of study manipulations, it was not possible to calculate separate meta-analyses. In addition, individual effect sizes for all experimental studies are reported in Appendix C.

Risk of bias

For experimental studies with randomised group allocation, risk of bias was assessed using the Cochrane Risk of Bias 2.0 tool (42). Studies using a within-subjects design were appraised using the checklist for cross-over studies by Ding et al. (43). For observational studies, risk of bias was assessed using the JBI Checklist for Analytical Cross Sectional Studies (44). Furthermore, a potential publication bias was investigated using funnel plots and Egger's test for funnel plot asymmetry (45). Funnel plots were adjusted using the trim and fill method (46) using metafor 3.0-2 (47) in R Studio 1.1.456/ R version 4.0.3.

Deviations from the protocol

Since truly objective digital assessment of consumption behaviours in daily life including smoking, alcohol consumption, and food intake is still in its early stages (21, 48), it was decided prior to screening titles and abstracts that Ecological Momentary Assessment (EMA) (18) of consumption behaviours would be considered as long as participants were asked to record the occasion immediately before, during or after consumption, preferably using objective markers such as a photo (21) which minimises recall bias (18).

Results

Literature search

A total of 11,723 individual records were screened. After 11,662 were excluded when screening titles and abstracts, 61 full texts were screened for eligibility. An additional 6 records were identified through forward and backward citation searches. A total of 30 publications reporting on 31 studies were included (see Figure 1 for the flow of records). Two of these studies overlap in the reported data: Ullrich et al. (49) extends the data reported in Baumann et al. (50) by adding data from a second point of measurement.

Study and sample characteristics

The 30 publications were published between 2008 and 2021. Most publications stemmed from the US (23%, n = 7) (38, 39, 51-55), followed by the UK (17%, n = 5) (40, 56-59) and Australia (10%, n = 3) (60-62). See Table 1 for a summary of the study characteristics.

The majority of the 31 included studies focused on different aspects of physical activity (81%, *n* = 25) including steps/ walking (*n* = 15) (40, 52, 53, 56, 57, 61-69), moderate to vigorous physical activity (MVPA; *n* = 9) (49-51, 55, 58, 62, 64, 70, 71), light physical activity (LPA; *n* = 6) (49, 50, 62, 64, 70, 71) and activity counts (*n* = 7) (39, 51, 58, 67, 70, 72, 73). Four studies also investigated reactivity to measurement of sedentary behaviour (13%) (49, 50, 62, 70). The remaining six studies focused on consumption behaviours such as alcohol consumption (10%, *n* = 3) (54, 60, 74) and smoking (3%, *n* = 1) (38) as well as medication adherence (7%, *n* = 2) (59, 75). One study also assessed the number of non-alcoholic drinks (3%) (74).

The majority of studies focused on adults (55%, n = 17) (38, 40, 49, 50, 54, 56, 57, 59, 60, 62, 65, 67-69, 74, 75), while twelve studies (39%) included children or adolescents in various age ranges between 3 and 18 years (39, 52, 53, 58, 61, 63, 64, 66, 70-73). Two studies (7%) specifically compared children, adolescents and adults (51, 55).

Study designs

A range of study designs was used to investigate reactivity to digital measurement of health behaviour. The majority of studies (68%, *n* = 21) (49-52, 54, 55, 59, 60, 62-69, 71, 72, 74, 75) used observational within-subjects designs to test whether behaviour changed across the study period. Typically, a statistically significant change in behaviour was interpreted as measurement reactivity, e.g. a decline in steps or an increase in sedentary time, reflecting an initial elevation (or reduction) due to reactivity and a gradual return to pre-assessment levels. Studies either presented tests of linear effects across the study period, treating time as a continuous variable (e.g., Labhart et al. (74), Poulton et al. (60), Ullrich et al. (49)), or compared behaviour on the first study day to behaviour on a varying number of subsequent days (e.g., Davis and Loprinzi (51), Haegele et al. (64)).

The remaining ten studies compared measurement reactivity in at least two different conditions, either between (*n* = 4; (38, 58, 61, 70)) or within (*n* = 6; (39, 40, 53, 56, 57, 73)) participants. Typically, different types of devices were compared. Typically, studies hypothesised that unobtrusive recording would lead to little or no measurement reactivity, while being aware of the device's purpose or even being able to see the recorded data would increase healthy behaviours (e.g., physical activity) or decrease unhealthy behaviours (e.g., sedentary behaviour). Five experimental studies additionally tested whether behaviour changed across the study period, e.g. between different days or weeks of the study (39, 40, 58, 61, 73).

Types of devices compared in experimental studies

The ten experimental studies used a range of (manipulated) devices to study reactivity. The majority of studies used a device for which the actual use was concealed (39); e.g., Clemes and Parker (56) concealed the pedometer as a body posture monitor (see also Clemes and Deans (40), Clemes et al. (57), Vanhelst et al. (70)). In three of these studies, the use of the device was concealed in the first part of the study before being revealed for the second part of the study (40, 56, 57). In four studies, behaviour was compared between wearing a sealed (i.e., no data visible on the device) and an unsealed tracking device (39, 53, 56, 61). In three studies, participants were also asked to copy the feedback from the device into a diary (40, 56, 57). Two other studies provided participants with a concealed (73) or sealed (58) device for monitoring and provided some participants with a second device for which the purpose of recording physical activity was known and the data visible. Finally, one study compared low vs high frequency sampling conditions (38).

Measurement reactivity in digital assessment of health behaviour

Physical activity

For seven experimental studies (53, 56-58, 61, 70, 73), effect sizes for the comparison of measurement reactivity manipulations could be obtained. First, a random-effects (RE) meta-analysis was conducted combining all studies independent of the manipulation or physical activity indicator (see Figure 2), yielding a small and statistically significant pooled effect size for Cohen's $d = 0.27$, 95% CI [0.16; 0.39] (test for overall effect: $Z = 4.58$, $p < .001$). There was substantial heterogeneity as indicated by I^2 of 78% ($Tau^2 = 0.04$, $H^2 = 4.60$, $df = 17$, $p < .001$; The Cochrane Collaboration (76)). Due to this heterogeneity, results of the narrative synthesis are reported separately per physical activity indicator. Another separate meta-analysis was conducted for steps; it was not possible to conduct further meta-analyses due to the small number of experimental comparisons. Asymmetry was examined using a funnel plot (see Figure 3) and Egger's test (45, 77), which was not significant ($p = .976$), thus not providing evidence for publication bias. The funnel plot, adjusted using the trim and fill method (46) are presented in Figures 5, Appendix D. The effect size estimate remained unchanged after trim and fill.

Steps/ walking. The majority of studies targeting physical activity recorded participants' step counts or stepping/ walking time using pedometers and accelerometers. Ten of a total of 15 studies found evidence for reactivity to the measurement. Evidence stems from both observational (62, 65-68) and experimental designs (40, 56, 57, 61). Four observational (52, 63, 64, 69) and one experimental study (53), on the other hand, did not find evidence for measurement reactivity. One observational study additionally reported evidence for reactivity to the measurement of stepping time, but not for standing time (62). A random-effects meta-analysis of the four experimental studies (53, 56, 57, 61) on steps (see Figure 3) yielded a statistically significant small to medium pooled effect size for Cohen's $d = 0.30$, 95% CI

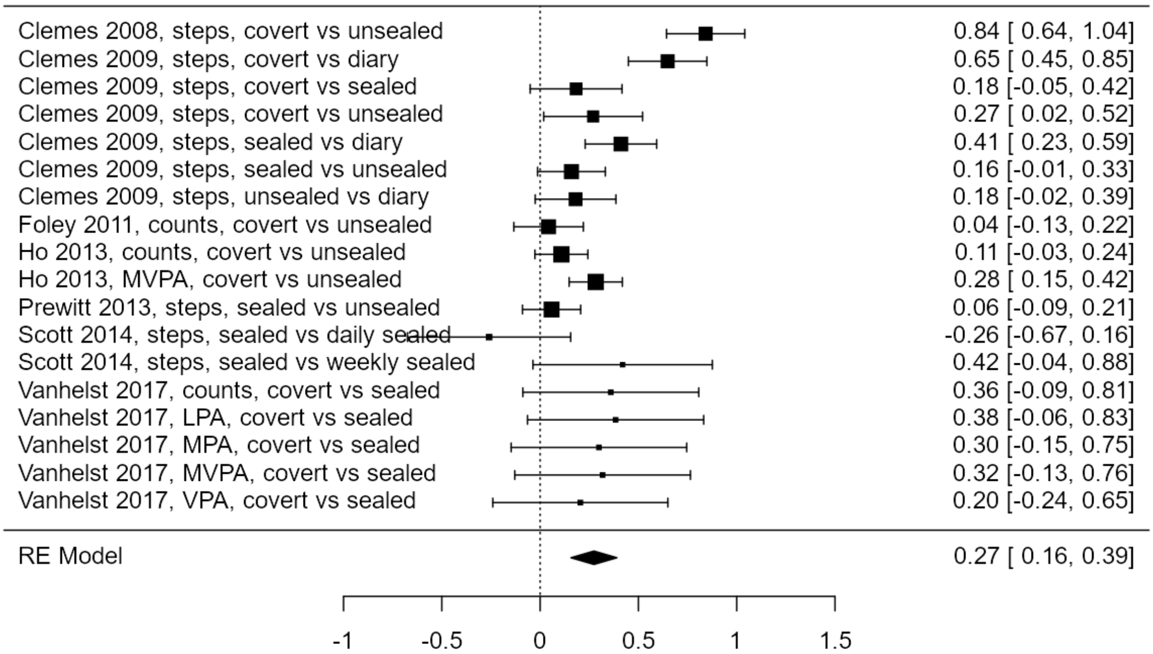

Figure 2. Forest plot of experimental studies with all physical activity outcomes. Effects on the right side of the dashed line indicate that manipulations that were hypothesised to increase measurement reactivity did indeed increase physical activity. Effects on the left side of the dashed line indicate the opposite effect. RE = random effects.

[0.12; 0.49] (test for overall effect: Z = 3.20, $p = 0.001$). Again, heterogeneity was substantial ($I^2 = 86\%$, $Tau^2 = 0.07$, $H^2 = 7.27$, $df = 9$, $p < .001$; The Cochrane Collaboration (76)). Asymmetry was investigated using a funnel plot (see Figure 3) and Egger's test (45, 77), which was not significant ($p = .475$), thus not providing evidence for publication bias. The funnel plot adjusted using the trim and fill method (46) is presented in Figure 6, Appendix D. Applying trim and fill increased the pooled effect size to Cohen's $d = 0.39$, 95% CI [0.19; 0.59], Z = 3.91, $p = .001$.

MVPA. A total of nine studies assessed moderate and/or vigorous physical activity using accelerometers. Two of these studies found a significant decrease in MVPA across the study period using an observational design, which is in line with the hypothesised effect (55,

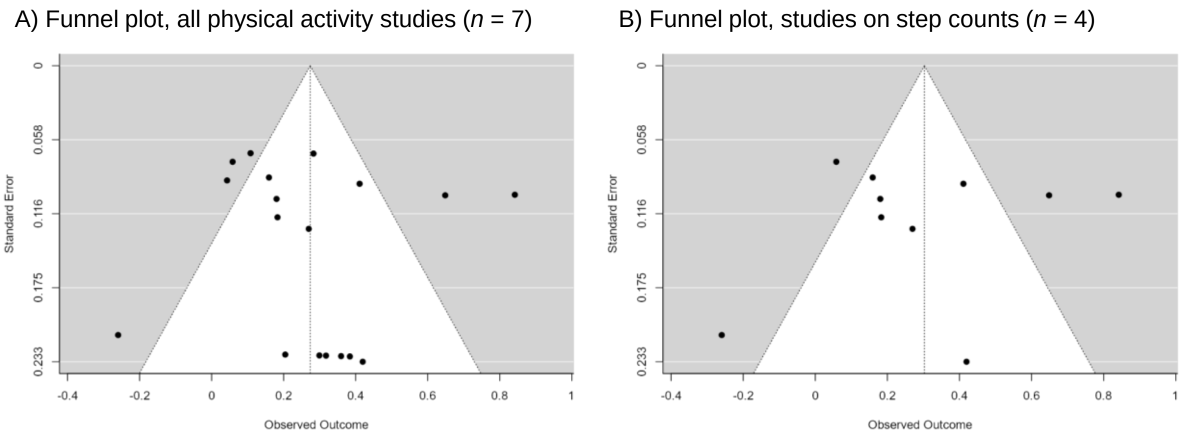

Figure 3. Funnel plots for all experimental studies investigating reactivity to measuring physical activity (panel A) and for all studies investigating reactivity to measuring step counts (panel B).

71). The remaining seven studies, including two experimental studies (58, 70), did not report measurement reactivity in their data (49-51, 58, 62, 64, 70).

LPA. Six studies assessed light physical activity using accelerometers. Three observational studies concluded that measurement reactivity occurred (49, 50, 71), while another observational study and one experimental study concluded that there was no reactivity to the assessment (62, 70). However, the non-significant effect in the experimental study was small to medium (Cohen's *d* of 0.38). One observational study reported changes in the data across the study period that were in line with the hypothesised measurement reactivity effect. The effect did not reach statistical significance but exceeded five percent change (64).

Counts. Seven studies assessed physical activity using activity counts provided by accelerometers. Two observational studies reported a decline in activity counts over time which was interpreted as measurement reactivity (67, 72). Three experimental studies also reported changes in activity counts depending on the condition; participants recorded higher activity counts if the counts were visible to them (39, 58, 73). The reported effects, however, were very small: Cohen's *d* ranged from 0.04 to 0.11. Foley et al. (73) reported an initial

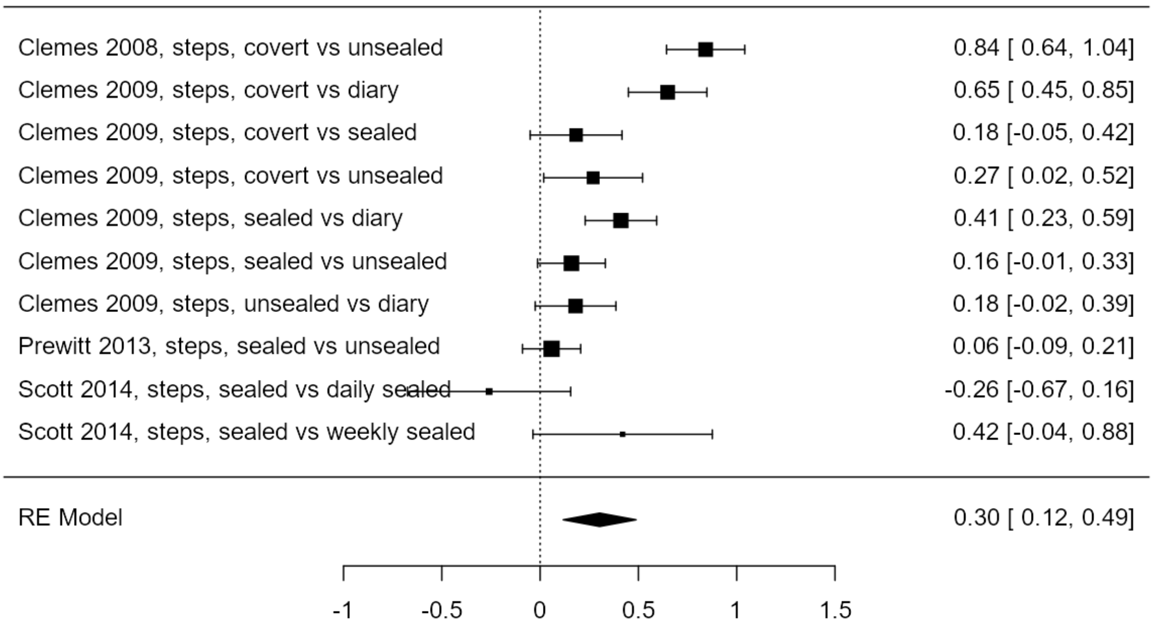

Figure 4. Forest plot of experimental studies focusing on steps. Effects on the right side of the dashed line indicate that manipulations that were hypothesised to increase measurement reactivity did indeed increase the number of recorded steps. Effects on the left side indicate the opposite direction. RE = random effects.

elevation of activity counts in the first thirty minutes of wearing the unsealed device, which attenuated afterwards. On the other hand, one observational (51) and one experimental study (Cohen’s d of 0.38; Vanhelst et al. (70)) concluded that measurement reactivity did not occur.

Sedentary behaviour

Four studies assessed sedentary time using accelerometers, which again reported mixed findings regarding the presence of measurement reactivity. Two observational studies reported a decline in sitting time across the study period, which was taken as an indicator for measurement reactivity (49, 50). Two other studies, one of which was experimental, concluded that measurement reactivity did not occur (62, 70). Although not statistically significant, the group difference in the experimental study indicated a small to medium effect of Cohen’s *d* = 0.42.

Alcohol consumption

Two out of three observational studies did not provide evidence of reactivity to digital measurement of the number of alcoholic drinks consumed using smartphone apps (60, 74). Yang et al. (54), on the other hand, reported a slight decrease in the number of drinks which levelled off at 25 days.

Consumption of non-alcoholic drinks

One observational study that focused on alcohol consumption also reported the number of non-alcoholic drinks consumed. This study did not find evidence for reactivity to the measurement (74).

Medication adherence

One observational study and observational data from a larger RCT on medication adherence both reported some measurement reactivity to using electronic monitoring systems; the effect only reached statistical significance in Cook et al. (75). They specifically point towards a drop in adherence between five and six weeks of recording. Sutton et al. (59) also report a small albeit nonsignificant decline in adherence across the study period.

Smoking

One study investigated reactivity to digital assessment of smoking cessation and tobacco abstinence using EMA devices (38). In this study, the prompting frequency was manipulated; some participants received one prompt per day to report smoking while other participants were prompted six times per day. The prompting frequency did not impact cessation or abstinence. The authors thus concluded that prompting frequency does not impact reactivity to the measurement.

Moderators

Seventeen studies included at least one moderator. There was substantial heterogeneity in the potential moderators of measurement reactivity explored in the included studies. No significant effects were found for whether activities in vs out of school were recorded (39, 70), the participants' Body-Mass Index, employment status (40), and the number of days on which a report was completed (74). Significant effects were found for the time point within a longitudinal study, with more reactivity occurring at the second compared to the first time point (49); the season in which the study was conducted with steeper increases in sedentary time and steeper decreases in LPA when data was collected in summer vs winter or spring; whether recording started on a weekday or at the weekend, with stronger declines in MVPA if measurement started on a weekday (50); the time of day when the recording took place, with measurement reactivity occurring earlier but not later in the day (73), visual disabilities, with children with visual disabilities showing an initial decline in MVPA while children and adults without visual disabilities showed an initial increase (55), the tendency to ruminate, with more pronounced reactivity in participants with a stronger tendency to ruminate (66); and hazardous drinking behaviour, with the effect only being present in participants who engaged in hazardous drinking behaviour, but not in participants who did not engage in hazardous drinking behaviour (60).

Indicators for the awareness of being measured

Six studies included moderators that reflected participants' awareness of being measured and their understanding of what participating in a study entails. Young children, for instance, may exhibit less social desirability bias due to lack of awareness. Accordingly, four studies included participants of different age groups ranging from children to adults (51, 53, 55, 72). Davis and Loprinzi (51) reported a stronger measurement reactivity effect in adults vs children, while Dössegger et al. (72) reported measurement reactivity to occur in children aged 7 years and older, but not in children between the age of 3 and 6. Zhu and Haegele (55), on the other hand, did not report systematic differences between children and adults. Similarly, Prewitt et al. (53) did not report differences between school children in grades 4, 5 or 6. Finally, Hilgenkamp et al. (65) compared effects in adults under and over the age of 65, reporting no significant differences.

Another moderator included in two studies was intellectual disability (65, 71), since it was hypothesised that people with intellectual disabilities may lack the understanding of the implications of measurement as part of a study. This hypothesis, however, was not confirmed: the strength of measurement reactivity did not differ between participants with low and high levels of intellectual disabilities.

In a similar vein, Prewitt et al. (53) tested for the moderating role of knowledge about pedometers. Again, no significant effect was found.

Gender

Six studies investigated gender differences in measurement reactivity. Five studies did not find significant differences between male and female participants (40, 51, 53, 56, 65). Ho et al. (58), on the other hand, found reactivity to measurement of activity counts in girls, but not in boys.

Risk of bias assessment

Risk of bias was assessed using three different tools, depending on the study design. Results are summarised in Tables 3 to 5 in Appendix D. For four studies using a randomised between-subjects design (38, 58, 61, 70), the Cochrane Risk of Bias 2.0 tool was used (42). All four studies were subject to significant risk of bias (see Table 2 for details), with two studies receiving the overall rating of some concerns (61, 70) and two receiving the overall rating of high risk of bias (38, 58). The high risk of bias arose from a lack of blinding, which was impossible due to the used assessment tool. Furthermore, since none of the four studies reported a pre-specified analysis plan, some risk of bias arose from the reported result.

Risk of bias of six studies using a within-subjects design (39, 40, 53, 56, 57, 73) was evaluated using the checklist for cross-over studies by Ding et al. (43). This checklist does not provide a summary evaluation. For all but one study, high risk of bias arose from non-randomised order of treatments (39, 40, 53, 56, 57); randomising the treatment order in these studies was not possible since they relied on participants being blinded to the use of the measurement device. Furthermore, no study specifically addressed potential carry-over effects of the treatments, resulting in unclear risk of bias. Similarly, for all six studies potential risk of bias arose from a lack of information on blinding participants and the research team. Finally, other potential sources of bias could not be assessed for all studies due to a lack of information.

Twenty-one observational studies were appraised using the JBI Checklist for Analytical Cross Sectional Studies (44), which again does not provide a summary evaluation. Some issues arose regarding defining inclusion criteria in six studies (49, 52, 63, 66, 71, 72) as well as regarding identifying and addressing confounding factors in nine (52, 54, 62, 66-69, 74)/ eight studies (52, 54, 62, 66-69), respectively.

Discussion

Summary of main findings

Digital in-the-moment assessment of health behaviour has become increasingly popular in recent years, mainly because it is believed to suffer less from typical shortcomings of self-report research such as recall bias (4). The present systematic review aimed to investigate the validity of health behaviour data collected in-the-moment with digital devices by synthesising the literature on reactivity to the measurement. The majority of identified studies focused on different aspects of physical activity. Overall, evidence for measurement reactivity was mixed. Effect sizes derived from experimental studies showed large heterogeneity and ranged from very small to large. However, the overall effect identified in the meta-analysis was small to medium (c.f. Cohen (35)), indicating that reactivity to the measurement of physical activity exists to some extent. The results may be useful when modelling research participation effects to quantitatively account for them in the data analysis without needing to formally assess measurement reactivity in individual studies (see Bendtsen and McCambridge (28) for recommendations).

The results of this systematic review indicate that measurement reactivity it not limited to self-reports in questionnaires but also extends to digitally assessed behavioural data. Indeed, confidence intervals of the meta-analysis suggest that reactivity for digitally assessed physical activity might be at least as strong, if not stronger, than for questionnaire-based assessments: previous meta-analyses produced pooled effect sizes (standardised mean difference, Hedge's g) of 0.19 to 0.21 (13, 15). This observation is in line with a previous meta-analysis that also reported stronger effects for objective compared to self-report measures, although this difference was not statistically significant (16). These findings challenge the assumption that objective measures of behaviour may lead to more ecologically valid conclusions than self-report measures; however, additional benefits of digital in-the-moment assessments such as reduced recall bias and the opportunity to study behaviour repeatedly and in daily life (78) remain.

Several studies included in this review suggest that measurement reactivity might be short-lived: Reactive effects are most likely to occur in the first hours to days of assessment and attenuate afterwards. It may thus be advised to exclude the data from the first days of use. For instance, Clemes and Deans (40) and Foote et al. (39) suggest to exclude data collected on physical activity from the first week of recording. Recommendations, however, may differ between behaviours; one study on medication adherence suggests to exclude data from the first six weeks of use (75), which may not always be feasible.

This systematic review highlights a lack of research on reactivity to the measurement of behaviours other than physical activity. Consumption behaviours such as medication adherence, alcohol consumption, smoking, and eating behaviour were rarely studied, which may be the case because they are more difficult to assess digitally than physical activity, for which passive tracking is well established. Most notably, although digital assessment of dietary intake including objective indicators such as photos is common in behavioural research (21), measurement reactivity has not yet been studied in this domain: Apart from one study that investigated the consumption of both alcoholic and non-alcoholic drinks during nights out, no studies on dietary intake were identified. Future research needs to address this gap to provide important information on the validity of digital assessment of dietary intake as well as other consumption behaviours.

Explanations for reactivity to measurement

Whether measurement reactivity occurs might depend on the indicator of the target behaviour, which varied substantially especially in the studies investigating physical activity. For instance, a large proportion of studies that investigated steps or activity counts as indicators for physical activity, which have been shown to correlate highly (79), reported reactive effects. Studies investigating moderate to vigorous physical activity, on the other hand, rarely reported reactive effects. Steps and activity counts may be more easily modifiable than moderate to vigorous physical activity since they require less effort and are easier to integrate in existing daily life activities (80). For instance, it may be easier to increase the number of steps by getting off the bus one stop early and walking the rest of the way than to schedule a formal exercise session at the gym on an already busy day. Only a small number of studies included in this review allowed for a direct comparison of different physical activity indicators by assessing several indicators in the same sample. The only study that included both steps and activity counts found reactive effects for both indicators (67). Moreover, Ho et al. (58) and Tinlin et al. (62) simultaneously assessed steps or activity counts and moderate to vigorous physical activity and reported reactivity to the measurement of the former, but not the latter indicator. Although they did not formally test this hypothesis, they support the notion that measurement reactivity is more likely to occur when the behaviour is easier to modify. On the other hand, Vanhelst et al. (70) reported reactive effects for neither indicator. Future research on measurement reactivity thus should explicitly take several indicators for the same behaviour into account that differ in the required effort to explicitly test this hypothesis.

This systematic review also investigated a large number of potential moderators of measurement reactivity. Based on the hypothesis that measurement reactivity is caused by forming beliefs about the study team's expectations that translate into behaviour via social desirability (25), several studies tested potential moderating effects of awareness of the measurement's purpose. Some studies experimentally manipulated awareness by concealing the purpose of the device (40, 56-58, 70, 73). Results from experimental studies were inconclusive, which could be explained by heterogeneity in the comparators (e.g., sealed or unsealed devices). Moreover, most studies used samples of less than 100 participants, which is too small to reliably detect the small to medium effects that the majority of studies reported. Other studies specifically recruited participants who were expected to be unaware of the implications of taking part in a study, such as children or people with intellectual disabilities (51, 53, 55, 65, 71, 72). However, it is unclear whether those subsamples were in fact aware of the implications of measurement in a research project since awareness was rarely assessed. A notable exception is Prewitt et al. (53), however, they investigated a narrow age range with school children in the 4^{th}, 5^{th} and 6^{th} grade. The results of this review regarding awareness as a potential underlying mechanism thus do not necessarily challenge this hypothesis, but rather underline the need for studying larger samples and choosing comparators carefully in future research.

Similarly, familiarity with tracking health behaviours might impact whether measurement reactivity occurs in the context of research. Especially tracking physical activity using wearables or smartphone apps has become popular in recent years (81, 82), so one might speculate that regular trackers may not experience increased awareness at the beginning of a study and may thus be less prone to measurement reactivity. This assumption, however, is yet to be tested in empirical research.

Some of the included studies suggest yet another explanation for measurement reactivity. For instance, several studies compared sealed vs unsealed devices. In both conditions, participants were aware of the measurement, but only when using an unsealed device, they also had access to the data that was collected. Having access to the data may serve as a reminder of study participation; however, it may also induce other cognitive processes such as reflection on the current behaviour (7). This may lead to detecting a discrepancy between the current and the desired behaviour, which in turn may trigger self-regulatory processes to adjust the behaviour (83, 84). Accordingly, when planning studies, researchers might need to avoid assessment tools such as accelerometers or fitness trackers that display the recorded data, or may need to take additional precautions so that the data are not visible, to reduce reactive effects (27). This may be especially important when using low-cost trackers that typically have displays (85). Although not all participants might necessarily have the intention to change behaviour when they enrol in the study, participants of health-related studies often are more health-conscious than the population average due to self-selection bias (also referred to as volunteer bias, see e.g. Haynes & Robinson (86) and Nuzzo (87) for discussions) and thus may also be more likely to have the intention to change their behaviour in line with recommendations. Thus, researchers might need to assess intention to change behaviour and introduce this variable as a covariate in the analysis to investigate potential intra-individual differences in measurement reactivity that may arise from differences in intention.

In a similar vein, Poulton et al. (60) investigated whether reactivity was stronger in participants who showed unhealthy alcohol consumption patterns; accordingly it could be hypothesised that measurement reactivity might be more pronounced in individuals who behave in a comparably unhealthy way and thus may see greater need for change (see also Barta et al. (7) for a summary). Indeed, only participants with hazardous drinking behaviour showed reactivity to measuring alcohol intake. Future research is needed to confirm these findings and to test their generalisability to other behaviours.

Limitations

It is important to address several shortcomings of this review and the included studies. Although a few studies with several hundred participants were included that were powered to detect small effects which would have been expected based on previous reviews on questionnaire-based research (13, 16), half of the included studies had less than 100 participants. Accordingly, uncertainty of the reported effects is large, as indicated by 95% confidence intervals of the pooled effect ranging from very small to medium. Especially studies with small samples might have missed small effects. Still, even small effects may have important implications for behavioural research (88), in which the mean effect size is small to medium (89), as well as clinical trials (27). Accordingly, researchers planning further studies on reactivity to digital measurement should account for potential small effects by recruiting sufficiently large samples. Moreover, future research needs to address shortcomings in study quality as outlined by quality appraisal tools; common issues included a lack of or missing information on blinding. Furthermore, no study used a pre-specified analysis plan; this can be solved through pre-registration.

Moreover, the studies included in the meta-analysis were heterogeneous in terms of a range of indicators for physical activity as well as study designs and compared devices. It was thus not possible to conduct meta-analyses separately for each physical activity indicator other than steps, and also the different manipulations could not be separated. More research is needed on systematically comparing different manipulations that may induce measurement reactivity to allow for the effects to be synthesised separately.

Two thirds of the studies included in this review were observational and they operationalized measurement reactivity as a particular pattern of change in behaviour across time. Observational study designs provide important insights into the temporal dynamics of measurement reactivity; however, changes in behaviour across time may also be induced by other influences such as participant fatigue and subsequent gaps in recording (90). Experimental designs are more robust, since they allow for the direct and controlled comparison of different conditions, but they may not always be feasible. For instance, it is possible to conceal the true purpose of certain devices to record physical activity; however, unobtrusive recording is typically not feasible or ethical for consumption behaviours (e.g., when using sensors that are applied to the participant's body, see Bell et al. (22) for an overview). Still, future research may provide important insights into key components of the measurement tool that amplify reactivity, e.g. by comparing measurement tools that differ in burden or complexity (27).

Lastly, due to the difficulties of measuring indicators of consumption behaviours (e.g., alcohol consumption, smoking, eating/ drinking) fully objectively, this review also included studies that investigated reactivity to digital recording of behaviour in-the-moment or close to consumption. This caveat is important to keep in mind when interpreting the findings.

However, through deviating from the pre-registered eligibility criteria for this review, a more complete picture of measurement reactivity research regarding digital assessments was obtained.

Conclusions

In summary, most research on reactivity to digital in-the-moment measurement of health behaviour focuses on physical activity. Measurement reactivity effects are generally small, but meaningful. The results extend a recently published list of study features that may indicate risk of bias due to measurement reactivity (27). For instance, measurement reactivity may be amplified when studying behaviours that require comparably little effort to change, such as the number of steps. Researchers using digital in-the-moment assessment tools might want to focus on moderate to vigorous physical activity, use assessment tools that do not provide participants access to the data, and use a run-in period of several days to a week to minimise reactive effects. More research is needed especially on potential reactivity to the assessment of consumption behaviours to be able to provide further guidance for researchers. Future studies should use experimental designs which enable different assessment methods to be compared and thus to identify the methods that induce least measurement reactivity. In this way, the validity of health behaviour research and thus the effectiveness of health promotion programmes can be improved.

References

1. Baumeister RF, Vohs KD, Funder DC. Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? Perspectives on Psychological Science. 2007;2(4):396-403.
2. Coughlin SS. Recall bias in epidemiologic studies. Journal of Clinical Epidemiology. 1990;43(1):87-91.
3. Winkler G, Döring A. Validation of a short qualitative food frequency list used in several German large scale surveys. Zeitschrift für Ernährungswissenschaft. 1998;37(3):234-41.
4. Thompson FE, Subar AF. Dietary assessment methodology. Nutrition in the Prevention and Treatment of Disease. 2017:5-48.
5. Grimm P. Social desirability bias. Wiley international encyclopedia of marketing. 2010.
6. Ram N, Brinberg M, Pincus AL, Conroy DE. The questionable ecological validity of ecological momentary assessment: Considerations for design and analysis. Research in Human Development. 2017;14(3):253-70.
7. Barta WD, Tennen H, Litt MD. Measurement reactivity in diary research. In: Mehl MR, Conner TS, editors. Handbook of Research Methods for Studying Daily life: The Guilford Press; 2012.
8. French DP, Sutton S. Reactivity of measurement in health psychology: how much of a problem is it? What can be done about it? British Journal of Health Psychology. 2010;15(3):453-68.
9. Sprott DE, Spangenberg ER, Block LG, Fitzsimons GJ, Morwitz VG, Williams P. The question–behavior effect: What we know and where we go from here. Social Influence. 2006;1(2):128-37.
10. Zajonc RB. Attitudinal effects of mere exposure. Journal of Personality and Social Psychology. 1968;9(2, Pt. 2):1-27.
11. Morwitz VG, Fitzsimons GJ. The mere-measurement effect: Why does measuring intentions change actual behavior? Journal of Consumer Psychology. 2004;14(1-2):64-74.
12. McCambridge J, Kypri K. Can simply answering research questions change behaviour? Systematic review and meta analyses of brief alcohol intervention trials. PloS one. 2011;6(10):e23748.
13. Miles LM, Rodrigues AM, Sniehotta FF, French DP. Asking questions changes health-related behavior: an updated systematic review and meta-analysis. Journal of Clinical Epidemiology. 2020;123:59-68.
14. Spangenberg ER, Kareklas I, Devezer B, Sprott DE. A meta-analytic synthesis of the question–behavior effect. Journal of Consumer Psychology. 2016;26(3):441-58.
15. Wilding S, Conner M, Sandberg T, Prestwich A, Lawton R, Wood C, et al. The question-behaviour effect: a theoretical and methodological review and meta-analysis. European Review of Social Psychology. 2016;27(1):196-230.
16. Wood C, Conner M, Miles E, Sandberg T, Taylor N, Godin G, et al. The impact of asking intention or self-prediction questions on subsequent behavior: a meta-analysis. Personality and Social Psychology Review. 2016;20(3):245-68.
17. Rodrigues AM, O'Brien N, French DP, Glidewell L, Sniehotta FF. The question–behavior effect: Genuine effect or spurious phenomenon? A systematic review of randomized controlled trials with meta-analyses. Health Psychology. 2015;34(1):61-78.
18. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annual Review of Clinical Psychology. 2008;4:1-32.

19. Müller AM, Maher CA, Vandelanotte C, Hingle M, Middelweerd A, Lopez ML, et al. Physical activity, sedentary behavior, and diet-related eHealth and mHealth research: bibliometric analysis. Journal of Medical Internet Research. 2018;20(4):e122.
20. Reichert M, Giurgiu M, Koch E, Wieland LM, Lautenbach S, Neubauer AB, et al. Ambulatory assessment for physical activity research: state of the science, best practices and future directions. Psychology of Sport and Exercise. 2020:101742.
21. König LM, Van Emmenis M, Nurmi J, Kassavou A, Sutton S. Characteristics of smartphone-based dietary assessment tools: A systematic review. PsyArXiv. 2021.
22. Bell BM, Alam R, Alshurafa N, Thomaz E, Mondol AS, de la Haye K, et al. Automatic, wearable-based, in-field eating detection approaches for public health research: a scoping review. NPJ digital medicine. 2020;3(1):1-14.
23. Degroote L, DeSmet A, De Bourdeaudhuij I, Van Dyck D, Crombez G. Content validity and methodological considerations in ecological momentary assessment studies on physical activity and sedentary behaviour: a systematic review. International Journal of Behavioral Nutrition and Physical Activity. 2020;17(1):1-13.
24. Michie S, Wood CE, Johnston M, Abraham C, Francis J, Hardeman W. Behaviour change techniques: the development and evaluation of a taxonomic method for reporting and describing behaviour change interventions (a suite of five studies involving consensus methods, randomised controlled trials and analysis of qualitative data). Health Technology Assessment. 2015;19(99).
25. McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. Journal of Clinical Epidemiology. 2014;67(3):267-77.
26. Miles LM, Elbourne D, Farmer A, Gulliford M, Locock L, McCambridge J, et al. Bias due to MEasurement Reactions In Trials to improve health (MERIT): protocol for research to develop MRC guidance. Trials. 2018;19(1):1-8.
27. French DP, Miles LM, Elbourne D, Farmer A, Gulliford M, Locock L, et al. Reducing bias in trials due to reactions to measurement: experts produced recommendations informed by evidence. Journal of Clinical Epidemiology. 2021.
28. Bendtsen M, McCambridge J. Causal models accounted for research participation effects when estimating effects in a behavioral intervention trial. Journal of Clinical Epidemiology. 2021;136:77-83.
29. French DP, Miles L, Elbourne D, Farmer A, Gulliford M, Locock L, et al. Reducing bias in trials from reactions to measurement: The MERIT study including developmental work and expert workshop. Health Technology Assessment. 2021.
30. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS medicine. 2009;6(7):e1000097.
31. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021;372.
32. Murray CJ, Abbafati C, Abbas KM, Abbasi M, Abbasi-Kangevari M, Abd-Allah F, et al. Five insights from the global burden of disease study 2019. The Lancet. 2020;396(10258):1135-59.
33. Hamilton WK. MAJOR: Meta-Analysis Jamovi R (Version 1.2.0). 2018.
34. The jamovi project. jamovi (Version 1.6.23). 2021.
35. Cohen J. A power primer. Psychological Bulletin. 1992;112(1):155-9.
36. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. Effect sizes based on means. Introduction to Meta-Analysis: Wiley; 2009. p. 21-32.

37. The Cochrane Collaboration. Table7.7.a: Formulae for combining groups. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions, Version 510: The Cochrane Collaboration; 2011.
38. McCarthy DE, Minami H, Yeh VM, Bold KW. An experimental investigation of reactivity to ecological momentary assessment frequency among adults trying to quit smoking. Addiction. 2015;110(10):1549-60.
39. Foote SJ, Wadsworth DD, Brock S, Hastie P, Cooper CK. The effect of a wrist worn accelerometer on children's in-school and out-of-school physical activity levels. Swedish Journal of Scientific Research. 2017;33(3):1-6.
40. Clemes SA, Deans NK. Presence and duration of reactivity to pedometers in adults. Medicine and Science in Sports and Exercise. 2012;44(6):1097-101.
41. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. Bmj. 2003;327(7414):557-60.
42. Sterne JA, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366.
43. Ding H, Hu GL, Zheng XY, Chen Q, Threapleton DE, Zhou ZH. The method quality of cross-over studies involved in Cochrane Systematic Reviews. PloS one. 2015;10(4):e0120519.
44. Moola S, Munn Z, Tufanaru C, Aromataris E, Sears K, Sfetcu R, et al. Chapter 7: Systematic reviews of etiology and risk In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis: JBI; 2020.
45. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. Bmj. 1997;315(7109):629-34.
46. Duval S, Tweedie R. Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000;56(2):455-63.
47. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of statistical software. 2010;36(3):1-48.
48. Höchsmann C, Martin CK. Review of the validity and feasibility of image-assisted methods for dietary assessment. International Journal of Obesity. 2020;44(12):2358-71.
49. Ullrich A, Baumann S, Voigt L, John U, Ulbricht S. Measurement Reactivity of Accelerometer-Based Sedentary Behavior and Physical Activity in 2 Assessment Periods. Journal of Physical Activity and Health. 2021;1(aop):1-7.
50. Baumann S, Groß S, Voigt L, Ullrich A, Weymar F, Schwaneberg T, et al. Pitfalls in accelerometer-based measurement of physical activity: The presence of reactivity in an adult population. Scandinavian Journal of Medicine & Science in Sports. 2018;28(3):1056-63.
51. Davis RE, Loprinzi PD. Examination of accelerometer reactivity among a population sample of children, adolescents, and adults. Journal of Physical Activity and Health. 2016;13(12):1325-32.
52. Ling J, King KM. Measuring physical activity of elementary school children with unsealed pedometers: compliance, reliability, and reactivity. Journal of Nursing Measurement. 2015;23(2):271-86.
53. Prewitt SL, Hannon JC, Brusseau TA. Children and pedometers: A study in reactivity and knowledge. International Journal of Exercise Science. 2013;6(3):230-5.
54. Yang C, Linas B, Kirk G, Bollinger R, Chang L, Chander G, et al. Feasibility and acceptability of smartphone-based ecological momentary assessment of alcohol use among African American men who have sex with men in Baltimore. JMIR mHealth and uHealth. 2015;3(2):e67.
55. Zhu X, Haegele JA. Reactivity to accelerometer measurement of children with visual impairments and their family members. Adapted Physical Activity Quarterly. 2019;36(4):492-500.

56. Clemes S, Parker RA. Increasing our understanding of reactivity to pedometers in adults. Medicine and Science in Sports and Exercise. 2009;41(3):674-80.
57. Clemes SA, Matchett N, Wane SL. Reactivity: an issue for short-term pedometer studies? British Journal of Sports Medicine. 2008;42(1):68-70.
58. Ho V, Simmons RK, Ridgway CL, van Sluijs EM, Bamber DJ, Goodyer IM, et al. Is wearing a pedometer associated with higher physical activity among adolescents? Preventive medicine. 2013;56(5):273-7.
59. Sutton S, Kinmonth A-L, Hardeman W, Hughes D, Boase S, Prevost AT, et al. Does electronic monitoring influence adherence to medication? Randomized controlled trial of measurement reactivity. Annals of Behavioral Medicine. 2014;48(3):293-9.
60. Poulton A, Pan J, Bruns Jr LR, Sinnott RO, Hester R. A smartphone app to assess alcohol consumption behavior: development, compliance, and reactivity. JMIR mHealth and uHealth. 2019;7(3):e11157.
61. Scott JJ, Morgan PJ, Plotnikoff RC, Trost SG, Lubans DR. Adolescent pedometer protocols: Examining reactivity, tampering and participants' perceptions. Journal of Sports Sciences. 2014;32(2):183-90.
62. Tinlin L, Fini N, Bernhardt J, Lewis LK, Olds T, English C. Best practice guidelines for the measurement of physical activity levels in stroke survivors: a secondary analysis of an observational study. International Journal of Rehabilitation Research. 2018;41(1):14-9.
63. Craig CL, Tudor-Locke C, Cragg S, Cameron C. Process and treatment of pedometer data collection for youth: the Canadian Physical Activity Levels among Youth study. Medicine and Science in Sports and Exercise. 2010;42(3):430-5.
64. Haegele JA, Zhu X, Bennett HJ. Brief Report: Reactivity to Accelerometer Measurement among Adolescents with Autism Spectrum Disorder. Journal of Autism and Developmental Disorders. 2020:1-5.
65. Hilgenkamp T, Van Wijck R, Evenhuis H. Measuring physical activity with pedometers in older adults with intellectual disability: reactivity and number of days. Intellectual and Developmental Disabilities. 2012;50(4):343-51.
66. Ling FC, Masters RS, McManus AM. Rehearsal and pedometer reactivity in children. Journal of Clinical Psychology. 2011;67(3):261-6.
67. Motl RW, Dlugonski D. Increasing physical activity in multiple sclerosis using a behavioral intervention. Behavioral Medicine. 2011;37(4):125-31.
68. Motl RW, McAuley E, Dlugonski D. Reactivity in baseline accelerometer data from a physical activity behavioral intervention. Health Psychology. 2012;31(2):172-5.
69. Klenk J, Peter RS, Rapp K, Dallmeier D, Rothenbacher D, Denkinger M, et al. 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. 2019;16(1):1-4.
70. Vanhelst J, Béghin L, Drumez E, Coopman S, Gottrand F. Awareness of wearing an accelerometer does not affect physical activity in youth. BMC Medical Research Methodology. 2017;17(1):1-6.
71. Zhu X, Haegele J, Wang D, Zhang L, Wu X. Reactivity to accelerometer measurement of youth with moderate and severe intellectual disabilities. Journal of Intellectual Disability Research. 2020;64(9):667-72.
72. Dössegger A, Ruch N, Jimmy G, Braun-Fahrländer C, Mäder U, Hänggi J, et al. Reactivity to accelerometer measurement of children and adolescents. Medicine and Science in Sports and Exercise. 2014;46(6):1140-6.
73. Foley JT, Beets MW, Cardinal BJ. Monitoring children's physical activity with pedometers: Reactivity revisited. Journal of Exercise Science & Fitness. 2011;9(2):82-6.
74. Labhart F, Tarsetti F, Bornet O, Santani D, Truong J, Landolt S, et al. Capturing drinking and nightlife behaviours and their social and physical context with a smartphone

application–investigation of users' experience and reactivity. Addiction Research & Theory. 2020;28(1):62-75.

75. Cook P, Schmiege S, McClean M, Aagaard L, Kahook M. Practical and analytic issues in the electronic assessment of adherence. Western Journal of Nursing Research. 2012;34(5):598-620.

76. The Cochrane Collaboration. 9.5.2 Identifying and measuring heterogeneity. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions2011.

77. Sterne JA, Egger M. Regression methods to detect publication and other bias in meta-analysis. In: Rothstein HR, Sutton AJ, Borenstein M, editors. Publication bias in meta-analysis: Prevention, assessment and adjustments. Chichester: Wiley; 2005. p. 99-110.

78. Trull TJ, Ebner-Priemer U. Ambulatory assessment. Annual Review of Clinical Psychology. 2013;9:151-76.

79. Sartini C, Wannamethee SG, Iliffe S, Morris RW, Ash S, Lennon L, et al. Diurnal patterns of objectively measured physical activity and sedentary behaviour in older men. BMC Public Health. 2015;15(1):1-13.

80. Cheval B, Boisgontier MP. The Theory of Effort Minimization in Physical Activity. Exercise and Sport Sciences Reviews. 2021;49(3):168-78.

81. König LM, Sproesser G, Schupp HT, Renner B. Describing the process of adopting nutrition and fitness apps: behavior stage model approach. JMIR mHealth and uHealth. 2018;6(3):e8261.

82. Pew Research Center. About one-in-five Americans use a smart watch or fitness tracker 2020 [Available from: https://www.pewresearch.org/fact-tank/2020/01/09/about-one-in-five-americans-use-a-smart-watch-or-fitness-tracker/.

83. Kanfer FH, Gaelick-Buys L. Self-management methods. In: Kanfer FH, Goldstein AP, editors. Helping people change: A textbook of methods: Pergamon Press; 1991. p. 305-60.

84. Carver CS, Scheier MF. Control theory: A useful conceptual framework for personality–social, clinical, and health psychology. Psychological Bulletin. 1982;92(1):111-35.

85. Degroote L, Hamerlinck G, Poels K, Maher C, Crombez G, De Bourdeaudhuij I, et al. Low-cost consumer-based trackers to measure physical activity and sleep duration among adults in free-living conditions: Validation study. JMIR mHealth and uHealth. 2020;8(5):e16674.

86. Haynes A, Robinson E. Who are we testing? Self-selection bias in laboratory-based eating behaviour studies. Appetite. 2019;141:104330.

87. Nuzzo J. Volunteer Bias and Female Participation in Exercise and Sports Science Research. Quest. 2021;73(1):82-101.

88. Götz FM, Gosling SD, Rentfrow P. Small Effects: The Indispensable Foundation for a Cumulative Psychological Science Perspectives on Psychological Science. 2021:1-11.

89. Cumming G, Calin-Jageman RJ. Introduction to the new statistics: Estimation, open science, and beyond: Routledge; 2017.

90. Ziesemer K, König LM, Boushey CJ, Villinger K, Wahl DR, Butscher S, et al. Occurrence of and reasons for "missing events" in mobile dietary assessments: results from three event-based ecological momentary assessment studies. JMIR mHealth and uHealth. 2020;8(10):e15430.

Table 1. Characteristics of the included studies.

Study	Country	Target behaviours	Assessment tool(s)	Study duration in days	Sample			Study design	Conditions		Moderators of measurement reactivity effect included in analysis	Conclusion [a]
					Age group(s)	Specific characteristics	N		Number	Description (if > 1)		
Baumann et al. (2018)[b]	Germany	PA (LPA, MVPA), SB (sitting time)	Accelerometer (ActiveGraph GT3X+)	7	adults	participants of a cardio-preventive health examination programme	160	Observational, within-subjects	1		Season (spring, summer, winter), first day of measurement (weekday vs weekend day)	(+/-) There is reactivity to measurement of LPA and SB. The effect is not significant for MVPA. Season moderates reactivity to measurement of LPA and SB with a steeper increase in SB and steeper decline in LPA in summer vs spring or winter. First day of measurement moderates reactivity to measurement of MVPA with stronger decline if measurement started on a weekday vs a weekend day.

Study	Country	Target behaviours	Assessment tool(s)	Study duration in days	Sample			Study design	Conditions		Moderators of measurement reactivity effect included in analysis	Conclusion [a]
					Age group(s)	Specific characteristics	N		Number	Description (if > 1)		
Clemes et al. (2008)	UK	PA (steps)	pedometer (New-Lifestyles NL-2000)	14	adults		50	Experimental, within-subjects	2	covert: pedometer concealed as a body posture monitor; unsealed + diary: use of pedometer known, in addition participant were asked to copy the daily step counts into a diary		(+) There is reactivity to measurement of steps.
Clemes and Parker (2009)	UK	PA (steps)	pedometer (New Lifestyles NL-1000)	28	adults		63	Experimental, within-subjects	4	covert: pedometer concealed as a body posture monitor; sealed: use of pedometer known, but display not visible; unsealed: use of pedometer known, display visible; diary: use of pedometer known,	gender	(+) There is reactivity to measurement of steps; effect is most pronounced when wearing an unsealed pedometer and recording step counts. Gender does not moderate the effect.

Study	Country	Target behaviours	Assessment tool(s)	Study duration in days	Sample			Study design	Conditions		Moderators of measurement reactivity effect included in analysis	Conclusion [a]
					Age group(s)	Specific characteristics	N		Number	Description (if > 1)		
										display visible, asked to copy daily step counts into diary		
Clemes and Deans (2012)	UK	PA (steps)	pedometer (New Lifestyles NL-1000)	21	adults		90	Experimental, within-subjects; also studies change in behaviour across time (study weeks)	2	covert: pedometer was concealed as a body posture monitor; overt: use of pedometer was announced, participants were asked to copy step counts into diary	BMI group, employment status (staff/ student), sex	(+) There is reactivity to measurement of steps; effect washes out after one week. BMI group, employment status and sex do not moderate the effect.
Cook et al. (2012)	NA	Medication adherence (% of prescribed doses taken)	Medication Event Monitoring System (MEMS)	84	adults	Patients with glaucoma	45	Observational, within-subjects	1			(+) There is reactivity to measurement of medication adherence; effect washes out after 5 weeks.
Craig et al. (2010)	Canada	PA (steps)	pedometer (SW-200)	7	Children and adolescents (5 to 19 years)		11477	Observational, within-subjects	1			(-) There is no reactivity to measurement of steps

Study	Country	Target behaviours	Assessment tool(s)	Study duration in days	Sample			Study design	Conditions		Moderators of measurement reactivity effect included in analysis	Conclusion [a]
					Age group(s)	Specific characteristics	N		Number	Description (if > 1)		
Davis and Loprinzi (2016)	US	PA (activity counts, MVPA)	accelerometer (ActiGraph 7164)	7	Children (6 to 11 years), adolescents (12-17 years), adults (18 to 85 years)		674	Observational, within-subjects	1		Age, gender, first day of monitoring (day of the week; weekday vs weekend day)	(-) There is not reactivity to measurement of MVPA or activity counts. There was some evidence that reactivity to measurement of activity counts was stronger in adults if the first day was a Monday.
Dössegger et al. (2014)	Switzerland	PA (activity counts)	accelerometer (ActiGraph models 7164, GT1M, GT3X)	7	Children, adolescents (3 years and older)		2081	Observational, within-subjects, data pooled from 8 studies	1		Age group (3-6 years, 7-11 years, ≥12 years), first day of monitoring (day of the week)	(+) There is reactivity to measurement of activity counts. The first day of monitoring moderated the effect; the effect was stronger for participants who started to monitor on a Wednesday compared to Sunday. Also age moderated the effect; measurement

Study	Country	Target behaviours	Assessment tool(s)	Study duration in days	Sample			Study design	Conditions		Moderators of measurement reactivity effect included in analysis	Conclusion [a]
					Age group(s)	Specific characteristics	N		Number	Description (if > 1)		
												reactivity was found starting from the age of 7.
Foley et al. (2011)	NA	PA (activity counts)	pedometer (Walk4Life 2525), accelerometer (Actiwatch)	20	Children (7-11 years)		32	Experimental, within-subjects; also studies changes in behaviour across time (within summer school blocks)	2	Treatment condition: participant wore both an accelerometer which was concealed as a watch to measure the time, and a pedometer which measured steps No treatment condition: Participants only wore the concealed accelerometer	Time of day (according to summer camp schedule)	(+/-) There is some evidence for reactivity to measurement of activity counts. This effect occurred in a warm-up period, but not for the overall assessment.
Foote et al. (2017)	US	PA (activity counts)	accelerometer (MOVABLE MOVband3)	16	children (10-12 years)		25	Experimental, within-subjects; also studies change in behaviour across time (study weeks)	2	Sealed accelerometer: participants were unable to see the activity counts (first study week); Unsealed accelerometer: participants	Activity in vs after school	(+) There is reactivity to measurement of activity counts, but only when wearing unsealed accelerometers: move counts were higher in first week of

Study	Country	Target behaviours	Assessment tool(s)	Study duration in days	Sample			Study design	Conditions		Moderators of measurement reactivity effect included in analysis	Conclusion [a]
					Age group(s)	Specific characteristics	N		Number	Description (if > 1)		
										were able to see the activity counts on the device (weeks 2-4)		wearing the unsealed accelerometer vs the second week of wearing the accelerometer, while there was no significant difference between the week in which the sealed accelerometer was worn vs the first week when the unsealed accelerometer was worn. Effects for reported separately for in school and out-of-school activities, but no test of an interaction was reported.
Haegele et al. (2020)	NA	PA (LPA, MVPA, step count/ wear time ratio)	accelerometer (ActiGraph GT3X)	7	Adolescents (13-18 years)	Autisim spectrum disorder	23	Observational, within-subjects	1			(+/-) There is no statistically significant effect of reactivity to the indicators of

Study	Country	Target behaviours	Assessment tool(s)	Study duration in days	Sample			Study design	Conditions		Moderators of measurement reactivity effect included in analysis	Conclusion [a]
					Age group(s)	Specific characteristics	N		Number	Description (if > 1)		
												PA, but change exceeds recommendation of 5% deviance on first or last day of measurement.
Hilgenkamp et al. (2012)	The Netherlands	PA (steps)	pedometer (Yamax Digi-Walker, NL-2000, NL-1000, NL-800)	14	adults (50 years and older)	with borderline to severe intellectual disability	135	Observational, within-subjects	1		Gender, age (below or above 65 years), level of intellectual disability, participants with/ without down syndrome	(-) There is no reactivity to measurement of steps. No moderators were identified.
Ho et al. (2013)	UK	PA (activity counts, MVPA)	accelerometer (Actiheart), pedometer (OMRON HJ-109)	4	adolescents (mean age 14.5 years)		892	Experimental, between-subjects; also studies change in behaviour across time (study days)	2	With pedometer: participants wore a pedometer in addition to an accelerometer; Without pedometer: participants only wore an accelerometer	Gender	(+/-) There is some evidence for reactivity to measurement of activity counts, but not for MVPA. Gender moderated the effect: viewing step counts of the pedometer was only associated with

Study	Country	Target behaviours	Assessment tool(s)	Study duration in days	Sample			Study design	Conditions		Moderators of measurement reactivity effect included in analysis	Conclusion [a]
					Age group(s)	Specific characteristics	N		Number	Description (if > 1)		
												increased activity counts in girls.
Klenk et al. (2019)	Germany	PA (walking time)	Accelerometer (activPAL)	5-7	Adults (65+ years)		1333	Observational, within-subjects	1			(-) There is no reactivity to measurement of walking duration.
Labhart et al. (2020)	Switzerland	Alcohol consumption (number of alcoholic drinks), diet (number of non-alcoholic drinks)	smartphone app (Youth@Night, Android)	49	adults (16-25 years)		241	Observational, within-subjects	1		Commitment level (assiduous, regular, and irregular participants)	(-) There is no reactivity to measurement of consumption of alcoholic and non-alcoholic drinks. Commitment level did not moderate the effect.
Ling et al. (2011)	Hong Kong	PA (steps)	pedometer (New Lifestyles NL-800)	21	children (9-12 years)		133	Observational, within-subjects	1		Rehearsal score	(+) There is reactivity to measurement of steps. The effect was more pronounced in high rehearsers.
Ling and King (2015)	US	PA (steps)	pedometer (Yamax SW-200)	7	children (mean age 9.25 years)		126	Observational, within-subjects	1			(-) There is no reactivity to measurement of steps.

Study	Country	Target behaviours	Assessment tool(s)	Study duration in days	Sample			Study design	Conditions		Moderators of measurement reactivity effect included in analysis	Conclusion [a]
					Age group(s)	Specific characteristics	N		Number	Description (if > 1)		
McCarthy et al. (2015)	US	Smoking (abstinence, cessation)	EMA device	28	Adults	smokers trying to quit who smoked at least 10 cigarettes per day	110	Experimental, between-subjects	2	high frequency condition: six prompts per day; low frequency condition: one prompt per day		(-) There is no reactivity to measurement of smoking when comparing high and low frequency recording.
Motl and Dlugonski (2011)	NA	PA (activity counts, steps)	accelerometer (ActiGraph 7164)	21	Adults	Multiple sclerosis patients	18	Observational, within-subjects	1			(+) There is reactivity to measurement of steps and activity counts.
Motl et al. (2012), Study 1	NA	PA (steps)	accelerometer (first 7 days: ActiGraog 7164; second 7 days: Omron HJ-720ITC)	14	Adults	Multiple sclerosis patients	18	Observational, within-subjects	1			(+) There is reactivity to measurement of steps.
Motl et al. (2012), Study 2	NA	PA (steps)	accelerometer (first 7 days: ActiGraog 7164; second 7 days: Omron HJ-720ITC)	14	Adults	Multiple sclerosis patients	20	Observational, within-subjects	1			(+) There is reactivity to measurement of steps.
Poulton et al. (2019)	Australia	Alcohol consumption (drinks per day)	smartphone app (CNLab-A, iOS and Android)	14	adults (16+ years)		671	Observational, within-subjects	1		Hazardous drinking (AUDIT)	(+) There is reactivity to measurement of alcohol consumption. The effect was only evident in

Study	Country	Target behaviours	Assessment tool(s)	Study duration in days	Sample			Study design	Conditions		Moderators of measurement reactivity effect included in analysis	Conclusion [a]
					Age group(s)	Specific characteristics	N		Number	Description (if > 1)		
												hazardous drinkers.
Prewitt et al. (2013)	US	PA (steps)	pedometer (Yamax Digi-Walker SW-200)	8	Children (4th to 6th grade)		109	Experimental, within-subjects	2	Sealed: steps counts were not visible; Unsealed: step counts were visible	Gender, grade, knowledge score quiz	(-) There is no reactivity to measurement of steps. The effect was not modulated by any of the moderators.
Scott et al. (2014)	Australia	PA (steps)	pedometer (Yamax Digi-Walker CW700), accelerometer (Actiigraph GT3X+)	7	adolescents (13-14 years)		96	Experimental, between-subjects; also studies change in behaviour across time (study days)	3	daily sealed pedometer: step counts were recorded daily by a research and a new sticker is put on the device; weekly sealed pedometer: display was sealed with a sticker; step counts were recorded by a researcher after the 7 days; unsealed: no sticker on the device		(+) There is reactivity to measurement of steps.
Sutton et al. (2014)	UK	Medication adherence	electronic medication-	56	adults	With Type 2 Diabetes	226	Observational, within-	1			(+/-) There may be reactivity to

Study	Country	Target behaviours	Assessment tool(s)	Study duration in days	Sample			Study design	Conditions		Moderators of measurement reactivity effect included in analysis	Conclusion [a]
					Age group(s)	Specific characteristics	N		Number	Description (if > 1)		
		(taking medication as prescribed)	monitoring device (TrackCap)					subjects; data from larger RCT comparing digital and non-digital assessment approaches for medication adherence				measurement of medication adherence, but effects were small and not significant.
Tinlin et al. (2018)	Australia	PA (LPA, MVPA, standing time, stepping time, steps), SB (sitting time)	accelerometers (activPAL3, Actigraph GT3X, Sensewear)	7	adults	Stroke history	32	Observational, within-subjects	1			(+/-) There is no reactivity to measurement of sitting time, standing time, LPA or MVPA. There is reactivity to measurement of steps and stepping time.
Ullrich et al. (2021) [b]	Germany	PA (LPA, MVPA), SB (sitting time)	accelerometer (ActiGraph GT3X+)	14	adults	participants of a cardio-preventive health examination programme	136	Observational, within-subjects	1		Time point in study (baseline vs at 12 months)	(+) There is reactivity to measurement of LPA and SB but not MVPA. The effect was moderated by time point in the study: measurement reactivity was stronger at

Study	Country	Target behaviours	Assessment tool(s)	Study duration in days	Sample			Study design	Conditions		Moderators of measurement reactivity effect included in analysis	Conclusion [a]
					Age group(s)	Specific characteristics	N		Number	Description (if > 1)		
												second time point compared to the first.
Vanhelst et al. (2017)	France	PA (activity counts, LPA time, MPA time, MVPA time, VPA time), SB (sedentary time)	accelerometer (ActiGraph GT3X)	4	adolescents (10-18 years)		78	Experimental, between-subjects	2	unaware of use of device: device concealed as body posture monitor; aware of use of device	Schooldays vs school-free days	(-) There is no reactivity to measurement of any of the included indicators for PA. There were no differences between schooldays and school-free days.
Yang et al. (2015)	US	Alcohol consumption (alcoholic drinks per day)	smartphone app (emocha, Android)	At least 24	adults	African American men who have sex with men	15	Observational, within-subjects	1			(+) There was a trends towards reactivity to measurement of alcohol consumption that flattened out after 25 days.
Zhu and Haegele (2019)	US	PA (MVPA)	accelerometer (ActiGraph GT3X)	4	children, adolescents (6-17 years), adults (parents)	children with visual impairments, their siblings and parents	66	Observational, within-subjects	1		Visual disability (yes/ no), age (parents vs children)	(+) There is reactivity to measurement of MVPA in children and adults without visual disabilities.

Study	Country	Target behaviours	Assessment tool(s)	Study duration in days	Sample			Study design	Conditions		Moderators of measurement reactivity effect included in analysis	Conclusion [a]
					Age group(s)	Specific characteristics	N		Number	Description (if > 1)		
												The direction of the effect differed: children without disabilities and their parents showed an initial elevation of MVPA, while children with visual disabilities showed an initial decline.
Zhu et al. (2020)	China	PA (LPA, MVPA)	accelerometer (ActiGraph GT3X)	7	Adolescents (mean age 13.9)	with moderate to severe intellectual disability	175	Observational, within-subjects	1		Severity of intellectual disability	(+) There is reactivity to measurement of LPA and MVPA for both disability groups.

Notes. [a] (+) measurement reactivity occurred, (-) measurement reactivity did not occur; [b] Partial overlap in the reported data. Abbreviations: EMA – Ecological Momentary Assessment, LPA – light physical activity, MEMS - Medication Event Monitoring System, MPA – moderate physical activity, MVPA – moderate to vigorous physical activity, PA – physical activity, SB – sedentary behaviour, VPA – vigorous physical activity

Appendix A – Search strategies

All searches were limited to 2008 onwards.

EMBASE

1. (((physical activity or pedomet* or acceleromet* or smok* or "tobacco use" or alcohol* or drink* or diet* or food or snack* or eating behav* or dental or tooth* or teeth* or medication* or tablet* or sedentary behav*) and (measure* or assess*) and reactiv*) not (c-reactive or diethyl* or reactive oxygen)).ab. or (((physical activity or pedomet* or acceleromet* or smok* or "tobacco use" or alcohol* or drink* or diet* or food or snack* or eating behav* or dental or tooth* or teeth* or medication* or tablet* or sedentary behav*) and (measure* or assess*) and reactiv*) not (c-reactive or diethyl* or reactive oxygen)).ti.

2. limit 1 to yr="2008 - 2020"

Pubmed (incl. MEDLINE)

(physical activity[Title/Abstract] OR pedomet*[Title/Abstract] OR acceleromet*[Title/Abstract] OR smok*[Title/Abstract] OR Tobacco Use[Title/Abstract] OR alcohol*[Title/Abstract] OR drink*[Title/Abstract] OR diet*[Title/Abstract] OR food[Title/Abstract] OR snack*[Title/Abstract] OR eating behav*[Title/Abstract] OR dental[Title/Abstract] OR tooth*[Title/Abstract] OR teeth*[Title/Abstract] OR medication*[Title/Abstract] OR tablet*[Title/Abstract] OR sedentary behav*)[Title/Abstract] AND (measure*[Title/Abstract] OR assess*)[Title/Abstract] AND reactiv* NOT (c-reactive[Title/Abstract] OR diethyl*[Title/Abstract] OR reactive oxygen)[Title/Abstract]

PsycInfo

TI ((physical activity OR pedomet* OR acceleromet* OR smok* OR Tobacco Use OR alcohol* OR drink* OR diet* OR food OR snack* OR eating behav* OR dental OR tooth* OR teeth* OR medication* OR tablet* OR sedentary behav*) AND (measure* OR assess*) AND reactiv* NOT (c-reactive OR diethyl* OR reactive oxygen)) OR AB ((physical activity OR pedomet* OR acceleromet* OR smok* OR Tobacco Use OR alcohol* OR drink* OR diet* OR food OR snack* OR eating behav* OR dental OR tooth* OR teeth* OR medication* OR tablet* OR sedentary behav*) AND (measure* OR assess*) AND reactiv* NOT (c-reactive OR diethyl* OR reactive oxygen))

Web of Science Core Collection

TI=((physical activity OR pedomet* OR acceleromet* OR smok* OR Tobacco Use OR alcohol* OR drink* OR diet* OR food OR snack* OR eating behav* OR dental OR tooth* OR teeth* OR medication* OR tablet* OR sedentary behav*) AND (measure* OR assess*) AND reactiv* NOT (c-reactive OR diethyl* OR reactive oxygen)) OR AB=((physical activity OR pedomet* OR acceleromet* OR smok* OR Tobacco Use OR alcohol* OR drink* OR diet* OR food OR snack* OR eating behav* OR dental OR tooth* OR teeth* OR medication* OR tablet* OR sedentary behav*) AND (measure* OR assess*) AND reactiv* NOT (c-reactive OR diethyl* OR reactive oxygen))

Appendix B – List of data extracted

Study information

- First author
- Year of publication
- Journal name
- Geographical setting
- Number of studies
- Target behaviour(s) and description
- Target group(s)
- Study duration in days
- Study design (within or between subjects)
- Description of study design
- Description of assessment tool(s)

Information on participants

- Number of participants
- Specific characteristics of the sample

Information on results

- Description of condition and comparator (if available)
- Type of analysis conducted
- Moderators
- Control variables
- Type of effect size reported
- Reported effect size
- Analysis statistically significant?
- M, SD of condition and comparator (if available)
- Overall conclusion regarding measurement reactivity

Appendix C

Table 2. Effect sizes for the experimental studies. Sufficient detail to compute effect sizes could be obtained for 7 of the 10 experimental studies.

Study	Behaviour	Comparison			Cohen's d [a]
		Between or within participants	Condition in which less reactivity was expected	Condition in which more reactivity was expected	
Clemes et al. (2008) [b]	Physical activity: steps	within	covert	unsealed	0.84
Clemes and Parker (2009) [b]	Physical activity: steps	within	covert	sealed	0.18
Clemes and Parker (2009) [b]	Physical activity: steps	within	covert	unsealed	0.27
Clemes and Parker (2009) [b]	Physical activity: steps	within	covert	diary	0.65
Clemes and Parker (2009) [b]	Physical activity: steps	within	sealed	unsealed	0.16
Clemes and Parker (2009) [b]	Physical activity: steps	within	sealed	diary	0.41

Study	Behaviour	Comparison			Cohen's d [a]
		Between or within participants	Condition in which less reactivity was expected	Condition in which more reactivity was expected	
Clemes and Parker (2009) [b]	Physical activity: steps	within	unsealed	diary	0.18
Clemes and Deans (2012)	Physical activity: steps	within	covert	diary	N/A
Foley et al. (2011) [c]	Physical activity: activity counts	within	covert	unsealed	0.04
Foote et al. (2017)	Physical activity: activity counts	within	sealed	unsealed	N/A
Ho et al. (2013) [d]	Physical activity: activity counts	between	covert	unsealed	0.11
Ho et al. (2013) [d]	Physical activity: MVPA	between	covert	unsealed	0.28
McCarthy et al. (2015)	Smoking	between	low frequency	high frequency	N/A
Prewitt et al. (2013) [b]	Physical activity: steps	within	sealed	unsealed	0.06

Study	Behaviour	Comparison			Cohen's d [a]
		Between or within participants	Condition in which less reactivity was expected	Condition in which more reactivity was expected	
Scott et al. (2014) [e]	Physical activity: steps	between	daily sealed	unsealed	-0.26
Scott et al. (2014) [e]	Physical activity: steps	between	weekly sealed	unsealed	0.42
Vanhelst et al. (2017) [e]	Physical activity: activity counts	between	covert	sealed	0.36
Vanhelst et al. (2017) [e]	Physical activity: LPA	between	covert	sealed	0.38
Vanhelst et al. (2017) [e]	Physical activity: MPA	between	covert	sealed	0.30
Vanhelst et al. (2017) [e]	Physical activity: VPA	between	covert	sealed	0.20
Vanhelst et al. (2017) [e]	Physical activity: MVPA	between	covert	sealed	0.32
Vanhelst et al. (2017) [e]	Sedentary behaviour	between	covert	sealed	0.42

Note. [a] Positive values indicate an effect in the assumed direction. [b] Effect size calculated based on original dataset retrieved from the authors based on Borenstein et al. (2009). [c] Effect size provided by the authors on request. [d] Separate results for boys and girls were combined as recommended in https://handbook-5-1.cochrane.org/chapter_7/table_7_7_a_formulae_for_combining_groups.htm. [e] Cohen's d was calculated based on means and standard deviations provided in the publication based on Borenstein et al. (2009).

Appendix D

Figure 5. Funnel plot for all experimental studies investigating reactivity to measuring physical activity that could be included in the meta-analysis, adjusted using the trim and fill method.

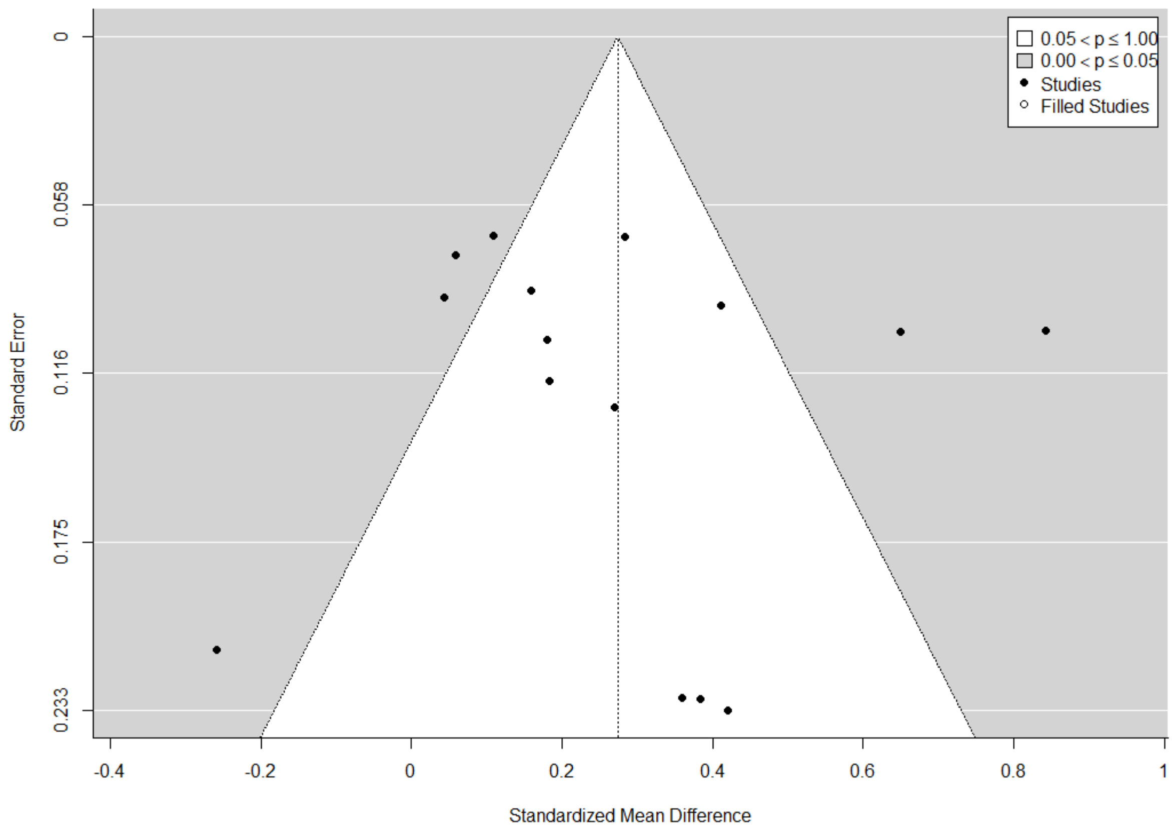

Figure 6. Funnel plot for all experimental studies investigating reactivity to measuring step counts, adjusted using the trim and fill method.

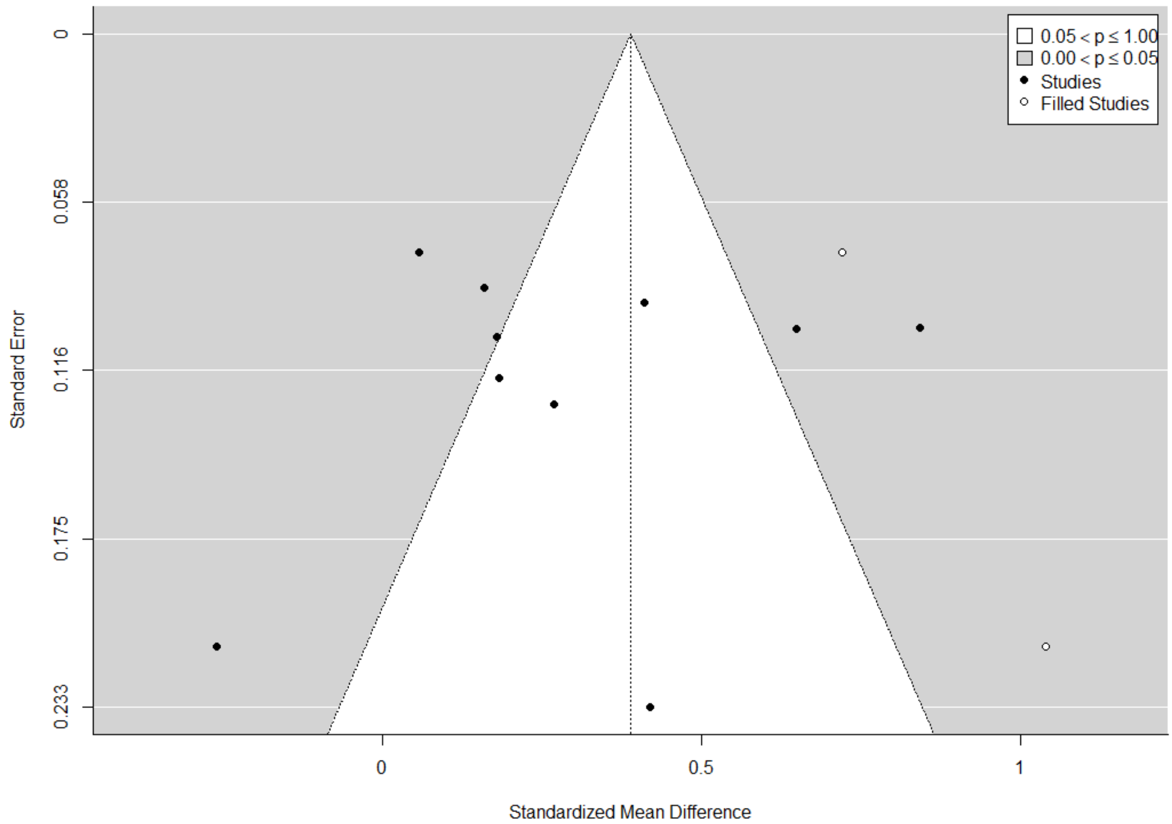

Table 3. Risk of bias assessment for randomised between-subject studies.

Study	Bias arising from the randomisation process	Bias due to deviations from the intended interventions (effect of assignment to intervention)	Bias due to missing outcome data	Bias in measurement of the outcome	Bias in selection of the reported result	Overall rating
Ho et al. (2013)	High	Some concerns	Low	Low	Some concerns	High
McCarthy et al. (2015)	High	Low	Low	Low	Some concerns	High
Scott et al. (2014)	Some concerns	Some concerns	Low	Low	Some concerns	Some concerns
Vanhelst et al. (2017)	Low	Some concerns	Low	Low	Some concerns	Some concerns

Table 4. Risk of bias assessment for studies using a within-subjects design.

	Appropriate cross-over design	Randomised treatment order	Carry-over effect	Unbiased data	Allocation concealment	Blinding	Incomplete outcome data	Selective outcome reporting	Other bias
Clemes et al. (2008)	Low	High	Unclear	Low	Low	Unclear	Unclear	Low	Unclear
Clemes and Parker (2009)	Low	High	Unclear	Low	Low	Unclear	Unclear	Low	Unclear
Clemes and Deans (2012)	Low	High	Unclear	Low	Low	Unclear	Low	Low	Unclear
Foley et al. (2011)	Low	Unclear	Unclear	Low	Unclear	Unclear	Low	Low	Unclear
Foote et al. (2017)	Low	High	Unclear	Low	Unclear	Unclear	Low	Low	Unclear
Prewitt et al. (2013)	Low	High	Unclear	Low	Unclear	Unclear	Low	Low	Unclear

Table 5. Risk of bias for observational studies.

Study	Were the criteria for inclusion in the sample clearly defined?	Were the study subjects and the setting described in detail?	Was the exposure measured in a valid and reliable way?	Were objective, standard criteria used for measurement of the condition?	Were confounding factors identified?	Were strategies to deal with confounding factors stated?	Were the outcomes measured in a valid and reliable way?	Was appropriate statistical analysis used?
Baumann et al. (2018)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cook et al. (2012)	Yes	No	Yes	Yes	Yes	Not clear	Not clear	Yes
Craig et al. (2010)	No	Yes	es	Yes	Yes	NA	Yes	Yes
Davis and Loprinzi (2016)	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Dössegger et al. (2014)	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Haegele et al. (2020)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hilgenkamp et al. (2012)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Labhart et al. (2020)	Yes	No	Not clear	No	No	Not clear	No	Not clear
Ling et al. (2011)	No	No	Yes	Yes	No	No	Yes	Yes
Ling and King (2015)	No	Yes	Yes	Yes	No	No	Yes	Yes
Klenk et al. (2019)	Yes	Yes	Yes	Yes	No	No	Yes	Yes
Motl and Dlugonski (2011)	Yes	Yes	Yes	Yes	No	No	Yes	Yes
Motl et al. (2012), Study 1	Yes	Yes	Yes	Yes	No	No	Yes	Yes

Motl et al. (2012), Study 2	Yes	Yes	Yes	Yes	No	No	Yes	Yes
Poulton et al. (2019)	Yes	Yes	Yes	No	Yes	Not clear	Yes	Yes
Sutton et al. (2014)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Tinlin et al. (2018)	Yes	Yes	Yes	Yes	No	No	Yes	Yes
Ullrich et al. (2021)	No	Yes	Yes	No	Yes	Yes	Yes	Yes
Yang et al. (2015)	Yes	Yes	Yes	Yes	No	No	Yes	Yes
Zhu and Haegele (2019)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Zhu et al. (2020)	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes